Fine-Gray competing risks model with high-dimensional covariates: estimation and Inference

Article *in* Electronic Journal of Statistics · July 2017

DOI: 10.1214/19-EJS1562

3 authors, including:

Jelena Bradic
University of California, San Diego
54 PUBLICATIONS **590** CITATIONS

SEE PROFILE

Fine-Gray Competing Risks Model with High-Dimensional Covariates: Estimation and Inference

Jue Hou[1], Jelena Bradic[1] and Ronghui Xu[1,2]
[1]Department of Mathematics,
[2]Department of Family Medicine and Public Health,
University of California San Diego, CA 92093

Abstract

The purpose of this paper is to construct confidence intervals for the regression coefficients in the Fine-Gray model for competing risks data with random censoring, where the number of covariates can be larger than the sample size. Despite strong motivation from biostatistics applications, high-dimensional Fine-Gray model has attracted relatively little attention among the methodological or theoretical literatures. We fill in this blank by proposing first a consistent regularized estimator and then the confidence intervals based on the one-step bias-correcting estimator. We are able to generalize the partial likelihood approach for the Fine-Gray model under random censoring despite many technical difficulties. We lay down a methodological and theoretical framework for the one-step bias-correcting estimator with the partial likelihood, which does not have independent and identically distributed entries. We also handle for our theory the approximation error from the inverse probability weighting (IPW), proposing novel concentration results for time dependent processes. In addition to the theoretical results and algorithms, we present extensive numerical experiments and an application to a study of non-cancer mortality among prostate cancer patients using the linked Medicare-SEER data.

Key words: p-values, survival analysis, high-dimensional inference, one-step estimator.

1 Introduction

In many applications, we want to use data to draw inferences about the effect of a covariate on a specific event in the presence of many risks competing for the same event: Examples include medical studies about the effect of a medical treatment on health outcomes of chronically ill patients, studies of the unemployment duration and transitions to employment and labor market programs, or evaluations of environmental determinants of child mortality, and studying internetwork competition risk "strategic gridlock," a study of how firms use alliances to respond to the alliance networks of their rivals. Historically, most datasets have been too small to meaningfully explore heterogeneity between different risk factors beyond considering cause-specific models only. Recently, however, there has been an explosion of experimental data sets where it is potentially feasible to develop estimates in full competing risks models.

High-dimensional regression has attracted increased interest in statistical analysis and has provided a useful tool in modern biomedical, ecological, astrophysical or economics data pertaining to setting where the number of parameters is greater than the number of samples (see Bühlmann and Van De Geer (2011) for an overview). Regularized methods (Fan and Li, 2001; Tibshirani, 1996) provide straightforward interpretation of resulting estimators, while allowing the number of covariates to be exponentially larger than the sample size. Considerable research effort has been devoted to developing regularized methods to handle various regression settings (Ravikumar *et al.*, 2010; Belloni and Chernozhukov, 2011; Obozinski *et al.*, 2011; Meinshausen and Bühlmann, 2006; Basu and Michailidis, 2015; Cho and Fryzlewicz, 2015) including those for time-to-event data (Sun *et al.* (2014); Bradic *et al.* (2011); Gaïffas and Guilloux (2012); Johnson (2008); Lemler (2013); Bradic and Song (2015); Huang *et al.* (2006); among others). However, regression has not been studied for the competing risks setting, a scenario frequently encountered in practice, with random censoring and high-dimensional covariates.

As an illustration project of how information contained in patients' electronic medical records can be harvested for the purposes of precision medicine, we consider the data set linking the Surveillance, Epidemiology and End Results (SEER) Program database of the National Cancer Institute with the federal health insurance program Medicare database for prostate cancer patients of age 65 or older. When restricted to patients diagnosed between 2004 and 2009 in the SEER-Medicare database, after excluding additional patients with missing clinical records, we have a total of 57,011 patients who have information available on 7 relevant clinical variables (age, PSA, Gleason score, AJCC stage, and AJCC stage T, N, M, respectively), 5 demographical variables (race, marital status, metro, registry and year of diagnosis), plus 8971 binary insurance claim codes. Until December 2013 (end of follow-up for this data) there were a total of 1,247 deaths due to cancer, and 5,221 deaths unrelated to cancer. The goal of this paper is to develop methodology for the Fine-Gray model with many more covariates than the number of events, which can be used to appropriately and flexibly evaluate the impact of risk factors on the non-cancer versus cancer mortality, as reflected in these clinical, demographical, and claim codes which indirectly describe events that occur in surgical procedures, hospitalization and outpatient activities. This understanding will then in turn aid in clinical decision making as to whether pursue aggressive cancer-directed therapy in the presence of pre-existing comorbidities.

There are at least three major challenges for addressing high-dimensional competing risks regression under the Fine-Gray model, which directly associates the risk factors with the cumulative incidence function of a particular cause. The structure of the score function related to the partial likelihood is a rather subtle issue with many of the unobserved factors ruining a simple martin-

gale representation. Shrinkage effects of the regularization methods add on a bias component that is non-ignorable when inference is of primary interest. Additionally, the structure of the sample information matrix prevents naive usage of Wald or Score type hypothesis testing methods, and basing theoretical analysis on the Hessian matrix renders problematic implementations. Attempts to tackle inference problems for the Fine-Gray regression model along this direction would also undesirably require implementation of bootstrap ideas. However, given the known problems of the bootstrap in high-dimensional setting this approach is no longer applicable. Development of high-dimensional inferential methods for competing risks data and the Fine-Gray model in particular, may have been hampered by these considerations.

In this paper we propose a natural and sensible formulation of inferential procedure for this high-dimensional competing risks regression. In the first step we formulate a l_1 regularized estimator of the high-dimensional parameter of interest and derive its finite-sample properties where the interplay between the sparsity and ambient dimension and the sample size can be directly seen. We note that our results are easily generalizable to a number of sparsity-inducing penalties but due to simplicity of presentation we present details only for the l_1 regularization. In the second step we formulate a bias-corrected estimator by formulating a new pragmatic estimator of the variance that allows broad dependence structures within the Fine-Gray model. This step compensates for the potential bias of the first estimator that arises due to variables that may be weakly correlated with the other risk scaues but are important due to their correlation with the risk of interest. We find that the second step estimator is effective at capturing strong signal as well as weak signals. This combination leads to an effective and simple-to-implement estimator in the Fine-Gray model with many features.

1.1 Setup and notation

For subject $i = 1, ..., n$ in a study, let T_i be the event time, with the event type or cause ϵ_i (we use the two words interchangeably in the following). Under the Fina-Gray model that we consider below, without loss of generality we assume that the event type of interest is '1', and we code all the other event types as '2' without further specifying them. In the presence of a potential right-censoring time C_i, the observed time is $X_i = T_i \wedge C_i$. The type of the event ϵ_i is also observed if $\delta_i = I(T_i \leq C_i) = 1$. Let $\mathbf{Z}_i(\cdot)$ be the vector of covariates that are possibly time-dependent. Assume that the observed data $\{(X_i, \delta_i, \delta_i\epsilon_i, \mathbf{Z}_i(\cdot))\}$ are independent and identically distributed (i.i.d.) for $i = 1, \ldots, n$.

Since the cumulative incidence function (CIF) is often the quantity of interest that can be estimated from data, Fine and Gray (1999) proposed a proportional subdistribution hazards model where the CIF

$$F_1(t|\mathbf{Z}_i(\cdot)) = \Pr(T_i \leq t, \epsilon_i = 1|\mathbf{Z}_i(\cdot)) = 1 - \exp\left(-\int_0^t e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)} h_0^1(u)du\right), \tag{1}$$

the p-dimensional coefficient $\boldsymbol{\beta}^o$ is the unknown parameter of interest, and $h_0^1(t)$ is the baseline subdistribution hazard. Under model (1) the corresponding subdistribution hazard $h_1(t|\mathbf{Z}_i(\cdot)) = h_0^1(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}$. Throughout the paper, we assume that there exists $s_o = |\text{supp}(\boldsymbol{\beta}^o)|$ for some $s_o \leq n$. Note that if we define an improper random variable $T_i^1 = T_i I(\epsilon_i = 1) + \infty I(\epsilon_i > 1)$, then the subdistribution hazard can be seen as the conditional hazard of T_i^1 given $\mathbf{Z}_i(\cdot)$. Fine and Gray (1999) proposed to estimate $\boldsymbol{\beta}$ based on the partial likelihood principle with complete data, i.e. when there is no censoring, and with censoring complete data. i.e. when the censoring times C_i's are always observed even if $C_i > T_i$. For the more general random censoring where $C_i > T_i$ is not

observed, inverse probability weighting (IPW) was used to obtain consistent estimating equations for $\boldsymbol{\beta}^o$.

We denote the counting process for type 1 event as $N_i^1(t) = I(T_i^1 \leq t)$ and its observed counterpart as $N_i^o(t) = I(\delta_i\epsilon_i = 1)I(X_i \leq t)$. We also denote the counting process for the censoring time as $N_i^c(t) = I(C_i \leq t)$. Let $Y_i(t) = 1 - N_i^1(t-)$ (note that this is not the 'at risk' indicator like under the classic Cox model), and $r_i(t) = I(C_i \geq T_i \wedge t)$. Note that $r_i(t)Y_i(t) = I(t \leq X_i) + I(t > X_i)I(\delta_i\epsilon_i > 1)$ is always observable, even though $Y_i(t)$ or $r_i(t)$ may not. Let $G(t) = \Pr(C_i \geq t)$ and let $\widehat{G}(\cdot)$ be the Kaplan-Meier estimator for $G(\cdot)$. Here we assume that C is independent of T, ϵ and $\mathbf{Z}$. Following the notation of Fine and Gray we call the IPW at-risk process:

$$\omega_i(t)Y_i(t) = r_i(t)Y_i(t)\frac{\widehat{G}(t)}{\widehat{G}(t \wedge X_i)}; \tag{2}$$

in other words, the weight for subject i is one if $t < X_i$, zero after being censored or failure due to cause 1, and $\widehat{G}(t)/\widehat{G}(X_i)$ after failure due to other causes. The modified log partial likelihood that gives rise to the weighted score function in Fine and Gray (1999) for $\boldsymbol{\beta}$ is

$$m(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left\{\boldsymbol{\beta}^\top\mathbf{Z}_i(t) - \log\left(\sum_{j=1}^{n}\omega_j(t)Y_j(t)e^{\boldsymbol{\beta}^\top\mathbf{Z}_j(t)}\right)\right\}dN_i^o(t), \tag{3}$$

where $t^* < \infty$ is the maximal follow-up time.

In the following for a vector $\mathbf{v}$, let $\mathbf{v}^{\otimes 0} = 1$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^\top$. We define for $l = 0, 1, 2$

$$\begin{aligned}\mathbf{s}^{(l)}(t,\boldsymbol{\beta}) &= \mathrm{E}\left\{G(t)/G(t \wedge X_i)r_i(t)Y_i(t)e^{\boldsymbol{\beta}^\top\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}\right\}, \quad \boldsymbol{\mu}(t) = \mathbf{s}^{(1)}(t,\boldsymbol{\beta}^o)/s^{(0)}(t,\boldsymbol{\beta}^o),\\ \mathbf{S}^{(l)}(t,\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n}\omega_i(t)Y_i(t)e^{\boldsymbol{\beta}^\top\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}, \quad \bar{\mathbf{Z}}(t,\boldsymbol{\beta}) = \mathbf{S}^{(1)}(t,\boldsymbol{\beta})/S^{(0)}(t,\boldsymbol{\beta}).\end{aligned} \tag{4}$$

1.2 Organization of the paper

This paper s organized as follows. In Section 2, we provide estimation and inference methodology developed for high-dimensional Fine-Gray model. Bounds for the prediction error of the Lasso estimator is presented in Section 3.1. We also discuss a related result whose rate matches those of linear models (see Theorem 1*). Section 3.2 studies the sampling distribution of a newly develop test statistics while allowing ultra high-dimensionality of the parameter space. We examine our regularized estimator and a one-step bias -correction estimator thorough Numerical Examples in Section 4 and a real data study in Section 5.

2 Estimation and inference for competing risks with more parameters than samples

2.1 Regularized estimator

A natural estimator to consider is a l_1-regularized estimator, where the particular loss function of interest would be the modified partial log-likelihood as defined in (3). That is, we consider

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\Big\{-m(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\Big\} \tag{5}$$

for a suitable choice of the tuning parameter $\lambda > 0$. Whenever possible we suppress λ in the notation above and use $\widehat{\boldsymbol{\beta}}$ to denote the l_1-regularized estimator. In this paper we quantify non-asymptotic oracle risk bound for the estimator above while allowing $p \gg n$ with a minimal set of assumptions. Theoretical studies of this kind is novel, since in the context of competing risks the martingale structures typically utilized in the high-dimensional work are ruined and new techniques needed to be developed. In particular, we show that the inverse probability weighting has a finite-sample effect that separates this model from the classic Cox model, for example (see comments after Theorem 1). We also establish that a certain tighter bound can be established whenever the hazard rate is bounded (see Theorem 1*). Finally, results presented therein can easily be broadened to any sparsity encouraging and convex penalty function.

2.2 One-step corrected estimator

For the purposes of constructing confidence intervals or testing significance of certain covariates, utilizing a naive regularized estimation as above is not appropriate; for example, construction of confidence intervals for those coefficients that have been shrunk to zero is impossible. On the other hand, firm guarantees of correct variable selection of $\widehat{\boldsymbol{\beta}}$ can only be established under restrictive set of assumptions including but not limited to the assumption of minimal signal strength of the true parameter $\boldsymbol{\beta}^o$ (Wasserman and Roeder, 2009; Fan and Lv, 2010; Meinshausen and Yu, 2009) which cannot be verified in practice. Therefore, it is of interest to develop inferential tools that do not depend on such assumptions and are yet able to provide theoretical guarantees of the quality of estimation and/or testing for example.

Inspired by the work of Zhang and Zhang (2014) and van de Geer *et al.* (2014), we propose the one-step bias-correction estimator

$$\widehat{\mathbf{b}} := \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}}), \tag{6}$$

where $\widehat{\boldsymbol{\beta}}$ is defined in (5), $\widehat{\boldsymbol{\Theta}}$ is an estimator of the "asymptotic" precision matrix $\boldsymbol{\Theta}$ to be defined below, and $\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}})$ is the score function, i.e. derivative of the modified log partial likelihood (3) evaluated at $\widehat{\boldsymbol{\beta}}$,

$$\dot{\mathbf{m}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t,\boldsymbol{\beta})\}dN_i^o(t).$$

The above construction of the one-step estimator is inspired by the first order Taylor expansion of $\dot{\mathbf{m}}(\cdot)$,

$$\dot{\mathbf{m}}(\boldsymbol{\beta}^o) \approx \dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}}) - \ddot{\mathbf{m}}(\boldsymbol{\beta}^o)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) \approx \ddot{\mathbf{m}}(\boldsymbol{\beta}^o)[\boldsymbol{\beta}^o - \{\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}})\}] = \ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\{\boldsymbol{\beta}^o - \widehat{\mathbf{b}}\}. \tag{7}$$

The notation $\approx$ in the above indicates that the equivalence is approximate with the higher order error terms omitted, and the negative Hessian $-\ddot{\mathbf{m}}(\boldsymbol{\beta}^o) \in \mathbb{R}^{p\times p}$ will have its limit denoted as $\boldsymbol{\Sigma}$. When $p \leq n$ an inverse of such a Hessian matrix would naturally be a good candidate for $\widehat{\boldsymbol{\Theta}}$ and with it $\boldsymbol{\Theta}^{-1} = \boldsymbol{\Sigma}$. However, when $p \geq n$ such an inverse does not necessarily exist. Therefore, we aim to find a good candidate matrix, $\widehat{\boldsymbol{\Theta}}$ such that $-\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\widehat{\boldsymbol{\Theta}} \approx \mathbb{I}_p$ with $\mathbb{I}_p$ denoting the $p \times p$ identity matrix. Although the construction here is inspired by the early works under the linear models, the specifics and theoretical analysis remain a challenge. In the following we will elucidate the construction of $\widehat{\boldsymbol{\Theta}}$.

2.3 Construction of the inverse Hessian matrix

We start by writing the negative Hessian of the modified log partial likelihood (3):

$$-\ddot{\mathbf{m}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left\{\frac{\mathbf{S}^{(2)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})} - \bar{\mathbf{Z}}(t,\boldsymbol{\beta})^{\otimes 2}\right\} dN_i^o(t). \tag{8}$$

We define

$$\boldsymbol{\Sigma} = \mathrm{E}\left[\int_0^{t^*}\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\}^{\otimes 2}\, dN_i^o(t)\right] = \mathrm{E}\left[\int_0^{t^*}\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\}\, dN_i^o(t)\right]^{\otimes 2}. \tag{9}$$

Under regularity conditions to be specified later, we have $\boldsymbol{\Sigma}$ as the "asymptotic negative Hessian" in the sense that the element-wise maximal norm $\|\boldsymbol{\Sigma}+\ddot{\mathbf{m}}(\boldsymbol{\beta})\|_{\max}$ converges to zero in probability. Our goal is to estimate its inverse $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_p)^\top$, where $\boldsymbol{\theta}_j$'s are the rows of $\boldsymbol{\Theta}$.

By (9), the positive semi-definite matrix $\boldsymbol{\Sigma}$ is also the second moment of the random vector

$$\boldsymbol{U}_i = \int_0^{t^*}\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\}\, dN_i^o(t) \tag{10}$$

with $\boldsymbol{\mu}(t)$ defined in (4). The expectation of $\boldsymbol{U}_i$ is zero,

$$\mathrm{E}(\boldsymbol{U}_i) = \mathrm{E}\left[\int_0^{t^*}\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\}\, Y_i(t) I(C_i\geq t) e^{\boldsymbol{\beta}^{t\top}\mathbf{Z}_i(o)} h_0^1(t)dt\right] = \mathbf{0}.$$

Hence, we may draw inspiration from the works on inverting high-dimensional variance-covariance matrix (Zhou *et al.*, 2011) to estimate our inverse $\boldsymbol{\Theta}$. Consider the minimizers of the expected loss functions

$$\boldsymbol{\gamma}_j^* = \underset{\boldsymbol{\gamma}_j\in\mathbb{R}^p}{\operatorname{argmin}}\,\mathrm{E}(U_j - \boldsymbol{U}_{-j}^\top\boldsymbol{\gamma}_j)^2, \quad \tau_j^2 = \mathrm{E}(U_j-\boldsymbol{U}_{-j}^\top\boldsymbol{\gamma}_j^*)^2, \tag{11}$$

where U_j is the jth element of $\boldsymbol{U}$, and $\boldsymbol{U}_{-j}$ is a $p-1$ dimensional vector created by dropping the jth element from $\boldsymbol{U}$. Note that τ_j^2 can also be alternatively written as

$$\mathrm{E}\{(U_j-\boldsymbol{U}_{-j}^\top\boldsymbol{\gamma}_j^*)U_j\} - \boldsymbol{\gamma}_j^{*\top}\mathrm{E}\{(U_j-\boldsymbol{U}_{-j}^\top\boldsymbol{\gamma}_j^*)\boldsymbol{U}_{-j}\}. \tag{12}$$

By the convexity of the target function $\mathrm{E}(U_j-\boldsymbol{U}_{-j}^\top\boldsymbol{\gamma}_j)^2$, $\boldsymbol{\gamma}_j^*$ must satisfy the first order Karush-Kuhn-Tucker conditions (KKT)

$$-\boldsymbol{\gamma}_j^{*\top}\mathrm{E}\{(U_j-\boldsymbol{U}_{-j}^\top\boldsymbol{\gamma}_j^*)\boldsymbol{U}_{-j}\} = 0. \tag{13}$$

Applying (13) to (12), we have

$$\tau_j^2 = \mathrm{E}\{(U_j-\boldsymbol{U}_{-j}^\top\boldsymbol{\gamma}_j^*)U_j\}.$$

We can then define a vector $\boldsymbol{\theta}_1 = (1,-\boldsymbol{\gamma}_1^{*\top})^\top/\tau_1^2$ that satisfies

$$\boldsymbol{\theta}_1^\top\boldsymbol{\Sigma} = \mathrm{E}\{(U_1-\boldsymbol{U}_{-1}^\top\boldsymbol{\gamma}_1^{*\top})\boldsymbol{U}\}/\,\mathrm{E}\{(U_1-\boldsymbol{U}_{-1}^\top\boldsymbol{\gamma}_1^*)U_1\} = (1,\mathbf{0}_{p-1}) = \mathbf{e}_1.$$

Without loss of generality, we may define $\boldsymbol{\theta}_j$ accordingly for $j=2,\ldots,p$, satisfying $\boldsymbol{\theta}_j^\top\boldsymbol{\Sigma} = \mathbf{e}_j$. The matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_p)^\top$ satisfies

$$\boldsymbol{\Theta}\boldsymbol{\Sigma} = (\mathbf{e}_1,\ldots,\mathbf{e}_p) = \mathbb{I}_p,$$

therefore it is the inverse of $\mathbf{\Sigma}$.

Using the empirical version of (11), we propose a consistent estimator for $\mathbf{\Theta}$. Due to the non-linearity of matrix inversion, the negative Hessian (8), in which we have the difference of two matrices inside an integral, is not easy to work with. Instead, we derive the sample version of (9) as the alternative form:

$$\widehat{\mathbf{\Sigma}} = n^{-1} \sum_{i=1}^{n} \int_0^{t^*} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t, \widehat{\boldsymbol{\beta}})\}^{\otimes 2} dN_i^o(t). \tag{14}$$

The advantage of $\widehat{\mathbf{\Sigma}}$ is that it can be written as the sample second moment $\widehat{\mathbf{\Sigma}} = n^{-1} \sum_{i=1}^{n} \widehat{\boldsymbol{U}}_i^{\otimes 2}$ where

$$\widehat{\boldsymbol{U}}_i = \int_0^{t^*} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t, \widehat{\boldsymbol{\beta}})\} dN_i^o(t). \tag{15}$$

We define

$$\Gamma_j(\boldsymbol{\gamma}_j, \widehat{\boldsymbol{\beta}}) = n^{-1} \sum_{i=1}^{n} \left(\widehat{U}_{i,j} - \widehat{\boldsymbol{U}}_{i,-j}^{\top} \boldsymbol{\gamma}_j \right)^2, \ j = 1, \ldots, p, \tag{16}$$

where $\widehat{U}_{i,j}$ is the jth element of $\widehat{\boldsymbol{U}}_i$, and $\widehat{\boldsymbol{U}}_{i,-j}$ is a $p-1$ dimensional vector obtained by dropping the jth element from $\widehat{\boldsymbol{U}}_i$. We then define the nodewise LASSO in our context to be

$$\widehat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}_j \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \{\Gamma_j(\boldsymbol{\gamma}_j, \widehat{\boldsymbol{\beta}}) + 2\lambda_j \|\boldsymbol{\gamma}_j\|_1\}, \quad \widehat{\tau}_j^2 = \Gamma_j(\widehat{\boldsymbol{\gamma}}_j, \widehat{\boldsymbol{\beta}}) + \lambda_j \|\widehat{\boldsymbol{\gamma}}_j\|_1. \tag{17}$$

Accordingly, we use $\widehat{\boldsymbol{\gamma}}_j$ and $\widehat{\tau}_j^2$ to construct

$$\widehat{\Theta}_{jk} = \begin{cases} -\widehat{\gamma}_{j,k}/(\widehat{\tau}_j^2), & k < j; \\ 1/(\widehat{\tau}_j^2), & k = j; \\ \widehat{\gamma}_{j,k-1}/(\widehat{\tau}_j^2), & k > j. \end{cases} \tag{18}$$

By the first order KKT condition, we have $(\widehat{\Theta}\widehat{\mathbf{\Sigma}})_{j,j} = 1$ and $|(\widehat{\Theta}\widehat{\mathbf{\Sigma}})_{j,k}| \leq \lambda_j$ for $j \neq k$. Choosing $\lambda_{\max} = \max_{j=1,\ldots,p} \lambda_j = o_p(1)$, we achieve that $\|\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}} - \mathbb{I}_p\|_{\max}$ goes to zero. The one-step estimator proposed in (6) with such $\widehat{\mathbf{\Theta}}$ hence converges to the true coefficient $\boldsymbol{\beta}^o$ approximately at the rate equivalent to $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$, as illustrated in (7).

Our proposed estimator is innovative in several aspects. Given the difficulty imposed by the model, we cannot make high-dimensional inference by simply inverting the $XX^{\top}$ with the covariate matrix X. The modified log partial likelihood (3) has dependent entries. The covariates $\mathbf{Z}_i(t)$ for $i = 1, \ldots, n$ are allowed to be time-dependent. Nevertheless, we identify for our model that the key element for the high-dimensional inference is each observation's contribution to the score, the $\boldsymbol{U}_i$'s. Our solution generalizes high-dimensional matrix inversion in a non-trivial way to complex models with non-standard likelihoods and weighting.

2.4 Confidence Intervals

To construct the confidence intervals for components of $\boldsymbol{\beta}^o$, we need the asymptotic distribution of $\widehat{\mathbf{b}}$. We first establish the asymptotic distribution of the score $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$. With $p > n$, we have to restrict the space in which we want to establish the asymptotic distribution. In general, it is impossible to establish convergence in distribution of $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ to a jointly Gaussian random variable

in $\mathbb{R}^p$ due to exploding dimensions. The asymptotic distribution for $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ is established in the following sense — for any $\mathbf{c} \in \mathbb{R}^p$ such that $\|\mathbf{c}\|_1 = 1$ we have

$$\sqrt{n}\mathbf{c}^\top \dot{\mathbf{m}}(\boldsymbol{\beta}^o) \xrightarrow{d} N(0, \mathbf{c}^\top \boldsymbol{\mathcal{V}} \mathbf{c}),$$

where $\boldsymbol{\mathcal{V}}$ is the variance-covariance matrix for $\sqrt{n}\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$. Obtaining the result above is technically challenging. As mentioned earlier apart from the high-dimensionality the modified log partial likelihood (3) has dependent summands. In addition, the IPW creates additional dependency through the Kaplan-Meier estimator. We construct the following estimator for $\boldsymbol{\mathcal{V}}$:

$$\widehat{\boldsymbol{\mathcal{V}}} = n^{-1} \sum_{i=1}^{n} (\widehat{\boldsymbol{\eta}}_i + \widehat{\boldsymbol{\psi}}_i)^{\otimes 2}, \tag{19}$$

where $\widehat{\eta}_i$ and $\widehat{\psi}_i$ are defined as follows:

$$\widehat{\boldsymbol{\eta}}_i = \int_0^{t^*} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t, \widehat{\boldsymbol{\beta}})\} \omega_i(t) d\widehat{M}_i^1(t), \tag{20}$$

$$\widehat{\boldsymbol{\psi}}_i = \int_0^{t^*} \frac{\widehat{\mathbf{q}}(t)}{\widehat{\pi}(t)} d\widehat{M}_i^c(t), \tag{21}$$

$$\widehat{\mathbf{q}}(t) = n^{-1} \sum_{i=1}^{n} I(t > X_i) \int_t^{t^*} \{\mathbf{Z}_i(u) - \bar{\mathbf{Z}}(u, \widehat{\boldsymbol{\beta}})\} \omega_i(u) d\widehat{M}_i^1(u), \tag{22}$$

$$\widehat{\pi}(t) = n^{-1} \sum_{i=1}^{n} I(X_i \geq t), \tag{23}$$

$$d\widehat{M}_i^1(t) = dN_i^o(t) - \frac{\omega_i(t) Y_i(t) e^{\widehat{\boldsymbol{\beta}}^\top \mathbf{Z}_i(t)}}{S^{(0)}(t, \widehat{\boldsymbol{\beta}})} n^{-1} \sum_{j=1}^{n} dN_j^o(t), \tag{24}$$

$$d\widehat{M}_i^c(t) = I(X_i \geq t) dN_i^c(t) - \frac{I(X_i \geq t)}{\widehat{\pi}(t)} n^{-1} \sum_{j=1}^{n} I(X_j \geq t) dN_j^c(t). \tag{25}$$

As illustrated in (7), we have $\sqrt{n}\mathbf{c}^\top(\widehat{\mathbf{b}} - \boldsymbol{\beta}^o)$ asymptotically equivalent to

$$\sqrt{n}\mathbf{c}^\top \boldsymbol{\Theta} \dot{\mathbf{m}}(\boldsymbol{\beta}^o) \xrightarrow{d} N(0, \mathbf{c}^\top \boldsymbol{\Theta} \boldsymbol{\mathcal{V}} \boldsymbol{\Theta}^\top \mathbf{c}).$$

We may now estimate the variance of $\sqrt{n}\mathbf{c}^\top(\widehat{\mathbf{b}} - \boldsymbol{\beta}^o)$ using a "sandwich" estimator $\mathbf{c}^\top \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\mathcal{V}}} \widehat{\boldsymbol{\Theta}}^\top \mathbf{c}$. Therefore a $(1-\alpha)100\%$ confidence interval for $\mathbf{c}^\top \boldsymbol{\beta}^o$ is

$$\left[\mathbf{c}^\top \widehat{\mathbf{b}} - \mathcal{Z}_{1-\alpha/2} \sqrt{\mathbf{c}^\top \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\mathcal{V}}} \widehat{\boldsymbol{\Theta}}^\top \mathbf{c}/n},\ \mathbf{c}^\top \widehat{\mathbf{b}} + \mathcal{Z}_{1-\alpha/2} \sqrt{\mathbf{c}^\top \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\mathcal{V}}} \widehat{\boldsymbol{\Theta}}^\top \mathbf{c}/n} \right] \tag{26}$$

with standard normal quantile $\mathcal{Z}_{1-\alpha/2}$.

Our proposed approach addresses various practical questions as special cases. First, we can construct confidence interval for a chosen coordinate β_j^o in $\boldsymbol{\beta}^o$. To that end, one needs to consider $\mathbf{c} = \mathbf{e}_j$, the j-th natural basis for $\mathbb{R}^p$ and apply the result (26). Generally, we can construct confidence interval for any linear contrasts $\mathbf{c}^\top \boldsymbol{\beta}^o$, potentially of any dimension. For example, we can have confidence intervals for the linear predictors $\mathbf{Z}^\top \boldsymbol{\beta}^o$ if the non-time-dependent covariate $\mathbf{Z}$

is also sparse so that we may assume $\|\mathbf{Z}\|_1$ to be bounded. As the dual problem, we may use the Wald test statistic

$$Z = \sqrt{n}(\mathbf{c}^\top \widehat{\mathbf{b}} - \theta_0)/\sqrt{\mathbf{c}^\top \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\mathcal{V}}}\widehat{\boldsymbol{\Theta}}^\top \mathbf{c}} \tag{27}$$

to test the hypothesis with $H_0 : \mathbf{c}^\top \boldsymbol{\beta}^o = \theta_0$.

3 Theory

Now we present the theory for the estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{b}}$ and the confidence intervals described in the previous section. First, we introduce some additional notations. Unlike the low-dimension situation in Fine and Gray (1999) where various empirical process results are applicable, none of the general results can be directly applied in high-dimension. This is a big challenge for our theory where the convergence of the empirical average of the time-dependent processes to their common expectation is needed at various places. We generalize some empirical process results while relying heavily on the martingale theory elsewhere.

The counting process martingales

$$M_i^1(t) = N_i^1(t) - \int_0^t Y_i(u) e^{\boldsymbol{\beta}^{o\top} \mathbf{Z}_i(u)} h_0^1(u) du \tag{28}$$

are essentially helpful tools in high-dimensions for establishing theory with dependent partial likelihoods. Unfortunately, the uncensored counting processes $\{N_i^1(t), i = 1, \ldots, n\}$ are not always observable. The observable counterpart $N_i^o(t)$ has no known martingale related to it under the observed filtration $\mathcal{F}_t = \sigma\{N_i^o(u), I(X_i \geq u), r_i(u) : u \leq t, i = 1, \ldots, n\}$. The Doob-Meyer compensator for the submartingale $N_i^o(t)$ under the observed filtration involves the nuisance distribution of $T_i|\epsilon_i > 1$. To utilize the martingale structure for our theory, we have to define the "censoring complete" filtration

$$\mathcal{F}_t^* = \sigma\{N_i^o(u), I(C_i \geq u), \mathbf{Z}_i(\cdot) : u \leq t, i = 1, \ldots, n\}, \tag{29}$$

on which we have a martingale $M_i^1(t)$ defined in (28),

$$\int_0^t I(C_i \geq t) dM_i^1(u) = N_i^o(t) - \int_0^t I(C_i \geq u) Y_i(u) e^{\boldsymbol{\beta}^{o\top} \mathbf{Z}_i(u)} h_0^1(u) du. \tag{30}$$

To relate the martingale (30) with our modified log partial likelihood $m(\boldsymbol{\beta})$, we define its proxy with $\mathcal{F}_t^*$ measurable integrand

$$\widetilde{m}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n \int_0^{t^*} \boldsymbol{\beta}^\top \mathbf{Z}_i(t) - \log \left(\sum_{j=1}^n I(C_j \geq t) Y_j(t) e^{\boldsymbol{\beta}^\top \mathbf{Z}_j(t)} \right) dN_i^o(t). \tag{31}$$

We define processes related to $\widetilde{m}(\boldsymbol{\beta})$ and its derivatives as

$$\widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n I(C_i \geq t) Y_i(t) e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes l}, \quad \widetilde{\mathbf{Z}}(t, \boldsymbol{\beta}) = \widetilde{\mathbf{S}}^{(1)}(t, \boldsymbol{\beta}) / \widetilde{S}^{(0)}(t, \boldsymbol{\beta}). \tag{32}$$

They can also be seen as proxies to the processes in (4). In fact, we can compute the following expectation by first conditioning on observed filtration

$$\mathrm{E}\left\{\widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta})\right\} = \mathrm{E}\left[\mathrm{E}\{I(C_i \geq t) Y_i(t) | \mathcal{F}_t\} e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes l}\right] = \mathrm{E}\left\{\widetilde{\omega}_i(t) Y_i(t) e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes 2}\right\},$$

where $\widetilde{\omega}_i(t) = r_i(t)G(t)/G(t \wedge X_i)$ is the weight with the true censoring distribution $G(\cdot)$. We denote their expectations as

$$\mathbf{s}^{(l)}(t, \boldsymbol{\beta}) = \mathrm{E}\left\{\widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta})\right\} = \mathrm{E}\left\{\widetilde{\omega}_i(t) Y_i(t) e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes 2}\right\}. \tag{33}$$

Our proxies precisely targets those weighted samples, as $\widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta})$ differs from $\mathbf{S}^{(l)}(t, \boldsymbol{\beta})$ only at those summands with observed type-2 events. We require below mild additional assumptions to control this approximation error.

Note that the Kaplan-Meier estimator for $G(t)$ can be written as

$$\widehat{G}(t) = \prod_{u \le t} \left(1 - \frac{dN_i^c(u)}{I(X_i \ge u)}\right).$$

To study the convergence of $\widehat{G}(t)$ to $G(t)$, we denote a martingale related to $N_i^c(t)$, the counting process of observed censoring. Let the censoring hazard be defined as $h^c(t) = -d\log(G(t))/dt$. Under the "censoring" filtration

$$\mathcal{F}_t = \sigma\{N_i^c(u), T_i, \epsilon_i, \mathbf{Z}_i(\cdot) : u \le t, i = 1, \ldots, n\}, \tag{34}$$

we have a martingale

$$M_i^c(t) = N_i^c(t) - \int_0^t I(C_i \ge u) h^c(u) du. \tag{35}$$

We use the integration-by-parts arguments (Murphy, 1994, the Helly-Bray argument on page 727) with random martingale measures, e.g. $dM_i^1(t)$, in our proof. In a rigorous sense, we should specify the element w in the probability space Ω for each $dM_i^1(t; w)$ when we apply integration-by-parts to the deterministic measure $dM_i^1(t; w)$ or solve integral equations involving $dM_i^1(t; w)$. The total variation of $M_i^1(t; w)$ is defined as

$$\bigvee_0^{t^*} M_i^1(t; w) = \sup_{k=1,2,\ldots} \sup_{0 \le t_1 < \cdots < t_k \le t^*} \sum_{j=2}^{n} |M_i^1(t_j; w) - M_i^1(t_{j-1}; w)|. \tag{36}$$

Since $M_i^1(t; w)$ can be decomposed into a nondecreasing counting process $N_i^1(t)$ minus another nondecreasing compensator $\int_0^t Y_i(u) e^{\boldsymbol{\beta}^{o\top} \mathbf{Z}_i(u)} h_0^1(u) du$, we have a bound for its total variation

$$\bigvee_0^{t^*} M_i^1(t; w) \le N_i^1(t^*) + \int_0^{t^*} Y_i(u) e^{\boldsymbol{\beta}^{o\top} \mathbf{Z}_i(u)} h_0^1(u) du. \tag{37}$$

Similar conclusion also applies to $M_i^c(t)$, i.e. we have a bound for its total variation

$$\bigvee_0^{t^*} M_i^c(t; w) \le N_i^c(t^*) + \int_0^{t^*} I(C_i \ge t) h^c(u) du. \tag{38}$$

As a convention, from hereon we suppress the w in the notation to keep it simple.

3.1 Oracle inequality

We first establish oracle inequality for the initial estimation error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1$ based on the following regularity conditions. All constants are assumed to be independent of n, p and s_o.

(C1) (Conditions on the design) Suppose that $\{(T_i, C_i, \epsilon_i, \mathbf{Z}_i(t)) : t \in [0,\infty)\}$ are i.i.d. with C_i independent of $(T_i, \epsilon_i, \mathbf{Z}_i)$. There exists a finite t^* such that $\Pr(C_i > t^*) = 0$. For any $t \in [0, t^*]$, $G(t) = I(C_i \geq t)$ is differentiable, and its hazard function $h^c(t) = -G'(t)/G(t) \leq K_c$. With probability equals one, the covariates satisfy

$$\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} \|\mathbf{Z}_i(t)\|_\infty \leq K/2. \tag{39}$$

There exist positive M and r_* such that

$$\inf_{t\in[0,t^*]} \mathrm{E}\left[I(C_i \geq t^*) I(t^* < T_i^1 < \infty) \min\{M, e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\}\right] > r_*. \tag{40}$$

(C2) (Continuity conditions for estimation) $\mathbf{Z}_i(t)$ may have jumps at $t_{i,1} < t_{i,2} < \cdots < t_{i,K_{zi}}$ with minimal gap between jumps bounded away from zero, $\min_{i=1,\ldots,n} \min_{1<k\leq K_{zi}} t_{i,k} - t_{i,k+1} \geq D$. Between two consecutive jumps, $\mathbf{Z}_i(t)$ has at most c_z elements Lipschitz continuous with Lipschitz constant L_z and the rest elements constant. The baseline CIF $F_1(t;\mathbf{0})$ is differentiable. The baseline subdistribution hazard $h_0^1(t) = -d\log\{F_1(t;\mathbf{0})\}/dt \in [h_*, K_h]$ over $(0, t^*)$ for $h_* > 0$ and $K_h < \infty$.

(C3) (Lower bound for restricted eigenvalue) For the M in (40), the smallest eigenvalue of matrix

$$\boldsymbol{\Sigma}(M) = \mathrm{E}\left\{\int_0^{t^*} \left(\mathbf{Z}(t) - \frac{\mathrm{E}\left[\mathbf{Z}(t)\{1 - F_1(t;\mathbf{Z})\}\min\{M, e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}(t)}\}\right]}{\mathrm{E}\left[\{1 - F_1(t;\mathbf{Z})\}\min\{M, e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}(t)}\}\right]}\right)^{\otimes 2} h_0^1(t)dt\right\}$$

is at least $\rho_* > 0$.

Remark 1. The conditions above are minimal in the sense that they appear in the oracle inequality under the Cox model (Huang *et al.*, 2013, (3.9) on page 1149; (4.5) and Theorem 4.1 on page 1154).

Remark 2. Due to missing censoring times among those with observed type-2 events, we have to make the additional assumptions to control the weighting errors. Although the weighted at-risk processes $\omega_i(t)$'s are asymptotically unbiased, the approximation errors in the tail $t \to \infty$ is poor for any finite n. To avoid unnecessary complications, we set the t^* as the final administrative censoring time, following the conventional design in the low-dimensional literature (Andersen and Gill, 1982). As a result, the partial likelihood (3) becomes an integral over the finite support $[0, t^*]$. We can let t^* go to infinity under more delicate assumptions by a martingale representation of approximation error (See Lemma B.8(ii) in supplement), but we decide not to replace the more straightforward assumptions here by the obscured ones that are harder to verify.

Remark 3. Condition (39) in (C1) is equivalent to the apparently weaker assumption (see for example Huang *et al.* (2013) equation (3.9)):

$$\sup_{1\leq i<j\leq n} \sup_{t\in[0,t^*]} \|\mathbf{Z}_i(t) - \mathbf{Z}_j(t)\|_\infty \leq K. \tag{41}$$

This can be seen by noting that the Cox model formulation is invariant when subtracting $\mathbf{Z}_i(t)$ by any deterministic $\boldsymbol{\zeta}(t)$.

Remark 4. Condition (40) in (C1) can be interpreted in two statements. First, the at-risk rate for type 1 events is bounded away from zero. Second, relative-risks arbitrarily close to zero needs to be truncated at a finite M; this is necessary in high-dimensions, in order to rule out the irregular situation where the non-zero expectation of the relative risks is dominated by a diminishing proportion of the excessively large relative risks. The same argument applies for (C3) in which a lower bound of the restricted eigenvalue of the negative Hessian (Bickel *et al.*, 2009) is defined.

Remark 5. Since the continuity condition (C2) may appear obscure, we offer some extra explanation. This has to do with the fact that the subjects with observed type 2 events remain indefinitely in the risk sets for type 1 events, as seen in the definitions of $\mathbf{S}^{(l)}(t, \boldsymbol{\beta}^o)$. These subject all have their $T_i^1 = \infty$, and under the proportional hazard model one would not expect their type 1 relative risks $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}$ to be excessively large. When the $\mathbf{Z}_i$'s are not time-dependent, we can use this fact to establish a slow growing rate of the maximal relative risks among those subjects. This is no longer the case with time-dependent $\mathbf{Z}_i(t)$'s, unless the linear predictor processes $\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)$ have certain continuity property. We propose (C2) taking into account likely practical scenarios, where the covariates are either measured at the baseline only, or otherwise at finitely many discrete time points. Note that the coordinate wise continuity in $\mathbf{Z}_i(t)$ is insufficient as p grows to infinity.

The concentration of the empirical average of the time-dependent processes around their expectation is one major challenge in establishing theory with high-dimensional covariates for survival data. We prove two widely applicable lemmas in the Appendix. Lemma A.1(i) produces concentration results for empirical average of processes with certain continuity property around their mean over an independently generated random grid. We use it to establish empirical bounds from the assumed population bound, and control the approximation errors. Lemma A.2(i) produces uniform concentration results for counting process martingales. We use it to establish sharper concentration results for martingales.

The regular deterministic cone-invertibility argument (Huang *et al.*, 2013; van de Geer, 2007; van de Geer and Bühlmann, 2009) remains valid under the Fine-Gray model. Let $\mathcal{O}$ be the indices set for non-zero elements in $\boldsymbol{\beta}^o$ and $\mathcal{O}_c$ be its compliment in $\{1, \ldots, p\}$. We define the cone set with $\xi > 1$

$$\mathcal{C}(\xi, \mathcal{O}) = \{\mathbf{b} \in \mathbb{R}^p : \|\mathbf{b}_{\mathcal{O}_c}\|_1 \leq \xi \|\mathbf{b}_{\mathcal{O}}\|_1\}, \tag{42}$$

and the compatibility factor with negative Hessian $-\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)$ on the cone set

$$\kappa\big(\xi, \mathcal{O}; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big) = \sup_{0 \neq \mathbf{b} \in \mathcal{C}(\xi, \mathcal{O})} \frac{\sqrt{s_o \mathbf{b}^\top \{-\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\mathbf{b}\}}}{\|\mathbf{b}_{\mathcal{O}}\|_1}. \tag{43}$$

On the event $\{\|\dot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_\infty < \lambda(\xi - 1)/(\xi + 1)\}$, the estimation error of LASSO estimator $\widehat{\boldsymbol{\beta}}$ defined in (5) has the bound

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \leq \frac{e^{\varsigma}(\xi + 1) s_o \lambda}{2\kappa\big(\xi, \mathcal{O}; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big)^2}, \tag{44}$$

where ς is the smaller solution to $\varsigma e^{-\varsigma} = K(\xi + 1) s_o \lambda / \{2\kappa\big(\xi, \mathcal{O}; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big)^2\}$.

The probabilistic condition of $\|\dot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_\infty$ decaying to zero, however, is not straightforward in the presence of both competing risks and censoring. The greatest challenge, as stated in the beginning of this section, is the lack of the martingale property in $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$. Even if we use its martingale proxy $\dot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$, the error in $\omega_i(t) - I(C_i \geq t)$ can be amplified in their product with the relative risks. In

the following we first show that the relative risks among subjects with observed type 2 events has sub-Gaussian tails. This is achieved through the argument that their CIF cannot be arbitrarily close to one. Otherwise, these subjects would have probability close to one experiencing type 1 event, contrary to the observed fact. As CIF is monotone increasing in relative risks, it is also unlikely to observe excessively large relative risks among the subjects with observed type 2 events. We then use Lemma A.1(i) to establish the concentration of $\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o) - \widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o)$ around zero across all observed type 1 event time. We state our result on the probabilistic condition for LASSO estimation error in the following lemma.

Lemma 1. *Denote* $K_{e,\varepsilon} = e^{c_z L_z \|\boldsymbol{\beta}^o\|_\infty D}\log(n/\varepsilon)/Dh_*$,

$$C^{(l)}_{n,p,\varepsilon} = \frac{K_{e,\varepsilon}K^l}{2^l}\left\{\frac{4M^2(1+K_c t^*)}{r_*^2}\sqrt{\frac{4\log(2/\varepsilon)}{n}} + \frac{4M^2K_c t^*}{r_*^2 n} + \sqrt{\frac{2\log(2np^l/\varepsilon)}{n}} + \frac{1}{n}\right\}, \tag{45}$$

and $\dot{C}(n,p,\varepsilon) = \left\{2C^{(1)}_{n,p,\varepsilon} + KC^{(0)}_{n,p,\varepsilon}\right\}/r_* + K\sqrt{2\log(2p/\varepsilon)/n}$. *Under Assumptions (C1) and (C2),*

$$\Pr\left(\|\dot{m}(\boldsymbol{\beta}^o)\|_\infty < \dot{C}(n,p,\varepsilon)\right) \geq 1 - e^{-nr_*^2/(2M^2)} - ne^{-n(r_*-2/n)^2/(8M^2)} - 5\epsilon.$$

This lemma then directly translates into a bound on the bias of estimation.

Theorem 1. *For* $\xi > 1$ *and a small* $\varepsilon > 0$, *set* $\lambda = \dot{C}(n,p,\varepsilon)(\xi-1)/(\xi+1)$ *with* $\dot{C}(n,p,\varepsilon)$ *from Lemma 1. Let* $C_\kappa > 0$ *satisfying* $\varsigma = 2K(\xi+1)s_o\lambda/(2C_\kappa)^2 \leq 1/e$, *and* $\eta \leq 1$ *be the smaller solution of* $\eta e^{-\eta} = \varsigma$. *Assume that* $n > -\log(\varepsilon/3)/(2p_*^2)$. *Under regularity conditions (C1) and (C2),*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 < \frac{e^\eta(\xi+1)s_o\lambda}{2C_\kappa^2}$$

occurs with probability no less than $\Pr\left(\kappa(\xi,\mathcal{O};\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)) > C_\kappa\right) - e^{-nr_*^2/(2M^2)} - ne^{-n(r_*-2/n)^2/(8M^2)} - 5\varepsilon$.

In Lemma 1, $K_{e,\varepsilon}$ comes from the ε-tail bound of the maximal relative risk among observed type 2 events. In contrary to its natural upper bound $e^{\|\boldsymbol{\beta}^o\|_1 K} \asymp e^{s_o}$, $K_{e,\varepsilon}$ is only of the order of $\log(n)$. $C^{(l)}_{n,p,\varepsilon}$ is the ε-tail bound of the approximation error

$$\sup_{k\in 1\ldots K_N} \|\mathbf{S}^{(l)}(T^1_{(k)},\boldsymbol{\beta}^o) - \widetilde{\mathbf{S}}^{(l)}(T^1_{(k)},\boldsymbol{\beta}^o)\|_{\max}$$

over the ordered observed type 1 event times $T^1_{(1)},\ldots,T^1_{(K_N)}$ with K_N being the total number of unique observed type 1 event times. Focusing on the leading terms, we have $C^{(0)}_{n,p,\varepsilon}$ of the order $\log(n)/\sqrt{n}$, and $C^{(1)}_{n,p,\varepsilon}$ of the order $\log(n)\sqrt{\log(p)/n}$. Hence, we obtain $\dot{C}(n,p,\varepsilon)$, the diminishing tail bound for $\|\dot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_\infty$, to be of the order $\log(n)\sqrt{\log(p)/n}$. Despite the complicated expression, our result differs from the rate $\sqrt{\log(p)/n}$ (established for the Cox or the linear model) only by a factor of $\log(n)$, which comes from the weights $\omega_i(t)$ among the terms of magnitude $e^{s_o} \gg n$. For the oracle inequality, we may directly assume the compatibility factor $\kappa\big(\xi,\mathcal{O};-\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big)$ to be bounded away from zero. Then, for the regularization parameter λ choosen to be of the order $\log(n)\sqrt{\log(p)/n}$ we obtain $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = O_p\left(s_o\log(n)\sqrt{\log(p)/n}\right)$.

The conclusion of the Theorem 1 involves the compatibility factor $\kappa\big(\xi,\mathcal{O};-\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big)$. To obtain $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1$ converging to zero in probability, we have to assume that $\Pr\left(\kappa(\xi,\mathcal{O};\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)) > C_\kappa\right)$

converges to zero for a sequence of C_κ bounded away from zero, as sample size n goes to infinity. Alternatively, we may achieve an equivalent conclusion under the condition (C3) which is used later for the asymptotic distribution. In the following lemma, we show that the negative Hessian has a lower bound converging to a positive definite matrix $\mathbf{\Sigma}(M)$ whose eigenvalues are assumed to be bounded away from zero by (C3). Thus, we obtain a lower bound of the restricted eigenvalue of the negative Hessian in the cone $\mathcal{C}(\xi, \mathcal{O})$. Using the connection between the compatibility factor and the restricted eigenvalue (van de Geer and Bühlmann, 2009), we show that $\kappa\big(\xi, \mathcal{O}; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big)$, the compatibility factor in the cone $\mathcal{C}(\xi, \mathcal{O})$, is bounded away from zero with probability tending to 1.

Lemma 2. *Denote*

$$\ddot{C}(n,p,\varepsilon) = \left\{2C^{(2)}_{n,p,\varepsilon} + 4KC^{(1)}_{n,p,\varepsilon} + (5/2)K^2C^{(0)}_{n,p,\varepsilon}\right\}/r_* + K^2\left\{(1+t^*K_h)\sqrt{2\log\big(p(p+1)/\varepsilon\big)/n} + (2/r_*)t^*K_h t^2_{n,p,\varepsilon}\right\},$$

where $t_{n,p,\varepsilon}$ is the solution of $p(p+1)\exp\{-nt^2_{n,p,\varepsilon}/(2+2t_{n,p,\varepsilon}/3)\} = \varepsilon/2.221$. Under Assumptions (C1)- (C3),

$$\Pr\left(\kappa\big(\xi, \mathcal{O}; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big) \geq \sqrt{\rho_* - s_o(\xi+1)\ddot{C}(n,p,\varepsilon)}\right) \geq 1 - 6\varepsilon.$$

Under the rate condition $s_o \log(n)\sqrt{\log(p)/n} = o(1)$, we obtain an asymptotically l_1-consistent regularized estimator $\widehat{\boldsymbol{\beta}}$. With $\widehat{\boldsymbol{\beta}}$ therefore being in a small neighborhood of the true parameter of interest $\boldsymbol{\beta}^o$, we are able to establish theory on statistical inference using the local quadratic approximation of $m(\boldsymbol{\beta})$ at $\boldsymbol{\beta}^o$.

3.2 Asymptotic normality for one-step estimator and honest coverage of confidence intervals

We establish the asymptotic normality for the one-step estimator $\widehat{\mathbf{b}}$ and honest coverage of the confidence intervals based on slightly stronger regularity conditions.

(D1) (Bound on linear predictors) The true linear predictors are uniformly bounded with probability one

$$\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\left|\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)\right| \leq K_b. \tag{46}$$

(D2) (Conditions on design) Suppose that $\{(T_i, C_i, \epsilon_i, \mathbf{Z}_i(t)) : t \in [0,\infty)\}$ are i.i.d., C_i is independent of $(T_i, \epsilon_i, \mathbf{Z}_i)$. There exists a finite t^* such that $\Pr(C_i > t^*) = 0$. For any $t \in [0, t^*]$, $G(t) = I(C_i \geq t)$ is differentiable, and its hazard function $h^c(t) = -G'(t)/G(t) \leq K_c$. With probability one, the covariates satisfy $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\|\mathbf{Z}_i(t)\|_\infty \leq K/2$. There exists $r_* > 0$, such that

$$\mathrm{E}\left[I(C_i \geq t^*)I(t^* \leq T_i^1 < \infty)\right] > r_*. \tag{47}$$

(D3) (Continuity conditions for inference) Each $\mathbf{Z}_i(t)$ is generated from random processes $\mathbf{d}_i^z(t)$, $\mathbf{\Delta}_i^z(t)$ and the counting process $N_i^z(t)$,

$$\mathbf{Z}_i(t) = \mathbf{Z}_i(0) + \int_0^t \mathbf{d}_i^z(u)du + \int_0^t \mathbf{\Delta}_i^z(u)dN_i^z(u).$$

$\boldsymbol{\beta}^{o\top}\mathbf{d}_i^z(t)$ is uniformly bounded between $\pm L_z$ and uniformly Lipschitz-L_{dz}. Each $\boldsymbol{\Delta}_i^z(t)$ is bounded in l^∞-norm by K in accordance with (D2), and $|\boldsymbol{\beta}^{o\top}\boldsymbol{\Delta}_i^z(t)| \leq 2K_b$ in accordance with (D1). The $N_i^z(t)$'s have a common cap on the number of jumps $K_z = o\left(\sqrt{n/(\log(p)\log(n))}\right)$ and a bounded intensity function $h^N(t) \leq L_N$. The baseline CIF $F_1(t;\mathbf{0})$ is differentiable. The baseline subdistribution hazard $h_0^1(t) = -d\log\{F_1(t;\mathbf{0})\}/dt$ exists and is bounded by K_h.

(D4) (Invertibility of negative Hessian) The smallest eigenvalue of the asymptotic Hessian matrix

$$\boldsymbol{\Sigma} = \mathrm{E}\left\{\int_0^{t^*} (\mathbf{Z}(t) - \boldsymbol{\mu}(t))^{\otimes 2}\, h_0^1(t)dt\right\},$$

with $\boldsymbol{\mu}(t) = \mathbf{s}^{(1)}(t,\boldsymbol{\beta}^o)/s^{(0)}(t,\boldsymbol{\beta}^o)$, is at least $\rho_* > 0$. The rows of the asymptotic precision matrix $(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_p)^\top = \boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ are sparse with sparsity constants $s_1,\ldots,s_p \leq s_{\max} < n$. Also, $\|\boldsymbol{\theta}_j/\Theta_{j,j}\|_1 \leq K_\gamma$.

(D5) (Rate condition for inference) The dimensions and sparsity parameters satisfy $s_o(s_{\max} + s_o)\log(p)/\sqrt{n} = o(1)$.

We make notably stronger assumptions in (D1) and (D3) than those required only for oracle inequality in the previous section. We explain their inevitability in the following discussion.

Like other works in high-dimensional inference beyond linear regression where normality is assumed, the restriction on the linear predictor $\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)$ becomes unavoidable (van de Geer *et al.*, 2014; Fang *et al.*, 2017). In our case, the asymptotic normality of $\sqrt{n}\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ depends fundamentally on the asymptotic tightness of $\sqrt{n}\dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$. As a necessary condition, the predictable quadratic variation under filtration $\mathcal{F}_t^*$ of the martingale $\sqrt{n},\dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$

$$\int_0^{t^*} n^{-1}\sum_{i=1}^n I(C_i \geq t)Y_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\{\mathbf{Z}_i(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}^{\otimes 2}h_0^1(t)dt, \tag{48}$$

must have a finite bound independent of the dimension of the covariates. This requires that the magnitude of the summands in (48) either be bounded or have light tails. Hence, we cannot allow the relative risk $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}$ to grow arbitrarily large. Additionally, we are able to achieve an improved estimation error for the initial estimator with (D1). The extra $\log(n)$ in the rate of Theorem 1 from bounding $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}$ with large probability is no longer necessary.

Theorem 1*. *Under (D1)-(D5), we can choose* $\lambda \asymp \sqrt{\log(np)/n}$ *and* $C_\kappa = \sqrt{p_*/2}$, *such that*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = O_p\left(s_o\sqrt{\log(p)/n}\right) = o_p(1).$$

In addition, we use Condition (D3) to derive a sharper $\sqrt{n}$-negligible bound for the approximation error between $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ and $\dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$ that is needed for the purposes of asymptotic normality. Naturally, the conclusion requires extra smoothness of $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}$'s compared to the Lipschitz continuity of (C2). We go one step further to make sure that their derivatives are also Lipschitz.

Finally, (D4) is a standard assumption for the validity of the nodewise penalized regressions (17). If we define the population versions of the nodewise components defined in (15)-(17),

$$\begin{aligned}
&\boldsymbol{U} = \int_0^{t^*}\{\mathbf{Z}(t) - \boldsymbol{\mu}(t)\}dN^o(t),\ \bar{\Gamma}_j(\boldsymbol{\gamma}) = \mathrm{E}\{U_j - \boldsymbol{U}_{i,-j}^\top\boldsymbol{\gamma}\}^2,\\
&\boldsymbol{\gamma}_j^* = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}}{\operatorname{argmin}}\,\bar{\Gamma}_j(\boldsymbol{\gamma}),\ \tau_j^2 = \bar{\Gamma}_j(\boldsymbol{\gamma}_j^*),
\end{aligned} \tag{49}$$

then the true parameters $\{\boldsymbol{\gamma}_j^*, \tau_j^2 : j = 1, \ldots, p\}$ uniquely define the inverse negative Hessian $\boldsymbol{\Theta}$ as described in Section 2.1. We prove this statement in the following Lemma.

Lemma 3. *Under (D4),* $\Theta j, j = 1/\tau_j^2$ *and* $\boldsymbol{\theta}_{j,-j}\tau_j^2 = \boldsymbol{\gamma}_j^*$. *Moreover,* $\|\boldsymbol{\gamma}_j^*\|_1 \leq K_\gamma$, $\tau_j^2 \geq \rho_*$ *and* $\|\boldsymbol{\Theta}\|_1 \leq K_\gamma/\rho_*$.

Lemma 3 also shows that $\|\boldsymbol{\gamma}_j^*\|_1$ and $\|\boldsymbol{\Theta}\|_1$ do not grow with p.

We use Lemma A.1(ii) and A.2(ii) to control the time-dependent processes for normality. Lemma A.1(ii) extends the uniform concentration over a countable grid in Lemma A.1(i) to the entire domain $[0, t^*]$, and this result is useful when we assess the approximation error between the two processes in an integral over absolutely continuous measures like $h_0^1(t)dt$. We apply Lemma A.2(ii) to bound $o_p\left(n^{-1/2}\right)$ errors involving martingale integrals with diminishing measurable integrands.

With all the preparation above, we start with the estimation error of the nodewise-LASSO. It is worth noting that the techniques from van de Geer and Bühlmann (2009) rely on i.i.d. entries in the linear regression model. Unlike those, the nodewise LASSO under our model (17) has dependent $\widehat{\boldsymbol{U}}_i$'s through the common $\bar{\mathbf{Z}}(t, \widehat{\boldsymbol{\beta}})$ in their definitions (15). Hence, our proof adopts the additional approximation of the dependent $\widehat{\boldsymbol{U}}_i$'s by their i.i.d. proxies. We eventually establish the error rates of $\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1$ and $|\widehat{\tau}_j^2 - \tau_j^2|$ in the following Lemma, which then leads to the error rate for $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1$ in Lemma 3.

Lemma 4. *Under (D1)-(D5), choosing* $\lambda_j \asymp s_o\sqrt{\log(p)/n}$, *we have* $\sup_j \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1 = O_p\left(s_o s_j \sqrt{\log(p)/n}\right)$ *and* $\sup_j |\widehat{\tau}_j^2 - \tau_j^2| = O_p(s_o s_j \sqrt{\log(p)/n})$. *Thus,* $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1 = O_p\left(s_o s_{\max}\sqrt{\log(p)/n}\right)$.

Remark 6. The total error of $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1$ is a product of the error from the initial estimator of the order s_o, the error from the nodewise LASSO of order $s_{\max}$ and the dimensions factor $\sqrt{\log(p)/n}$. Compared to the linear regression case (Zhang and Zhang, 2014; van de Geer *et al.*, 2014), our $\widehat{\boldsymbol{U}}$'s are affected by the estimation error of initial estimator $\widehat{\boldsymbol{\beta}}$. It therefore makes sense to have the extra s_o in our rate. Compared to the generalized linear model (GLM) case (van de Geer *et al.*, 2014), our $\widehat{\boldsymbol{U}}$'s are dependent with each other through $\bar{\mathbf{Z}}(t, \widehat{\boldsymbol{\beta}})$ involving the initial estimator. Thus, our error rate from two different sources takes the multiplicative form $s_o s_{\max}$ instead of the summation $s_o + s_{\max}$ for the GLM. In general, we consider our rate to be optimal under the our model.

Using Lemma 4, we can establish the approximation condition for $\widehat{\mathbf{b}}$ proposed in (7).

Lemma 5. *Under (D1)-(D5), the one-step estimator* $\widehat{\mathbf{b}}$ *satisfies the approximation condition*

$$\sqrt{n}\mathbf{c}^\top \left\{\boldsymbol{\Theta}\dot{\mathbf{m}}(\boldsymbol{\beta}^o) + \boldsymbol{\beta}^o - \widehat{\mathbf{b}}\right\} = O_p\left(s_o(s_{\max} + s_o)\log(p)/\sqrt{n}\right) = o_p(1)$$

for any $\mathbf{c}$ *such that* $\|\mathbf{c}\|_1 = 1$.

Next, we show the asymptotic normality of $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$. Define the asymptotic variance of $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ similar to that from Fine and Gray (1999) as

$$\boldsymbol{\mathcal{V}} = \mathrm{E}\{\boldsymbol{\eta}_i + \boldsymbol{\psi}_i\}^{\otimes 2}, \tag{50}$$

with

$$\boldsymbol{\eta}_i = \int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\widetilde{\omega}_i(t) dM_i^1(t), \tag{51}$$

$$\boldsymbol{\psi}_i = \int_0^{t^*} \int_0^{t^*} \frac{\mathbf{q}(t)}{\pi(t)} I(X_i \geq t) dM_i^c(t), \tag{52}$$

$$\mathbf{q}(t) = \mathrm{E}\left[I(t > X_i) \int_t^{t^*} \{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u) dM_i^1(u) \right], \tag{53}$$

$$\pi(t) = \Pr(X_i \geq u), \tag{54}$$

and $M_i^1(t)$, $M_i^c(t)$ as defined in (28) and (35).

Lemma 6. *Under conditions (D1)-(D5), for directional vector* $\mathbf{c} \in \mathbb{R}^p$ *with* $\|\mathbf{c}\|_1 = 1$ *and* $\mathbf{c}^\top \boldsymbol{\Theta}\boldsymbol{\mathcal{V}}\boldsymbol{\Theta}^\top \mathbf{c} \to \nu^2 \in (0, \infty)$, $\sqrt{n}\mathbf{c}^\top \boldsymbol{\Theta}\dot{\mathbf{m}}(\boldsymbol{\beta}^o)/\sqrt{\mathbf{c}^\top \boldsymbol{\Theta}\boldsymbol{\mathcal{V}}\boldsymbol{\Theta}^\top \mathbf{c}} \xrightarrow{d} N(0,1)$.

The proof uses the same approach as the initial low-dimensional result in Fine and Gray (1999). We approximate $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ by the sample average of i.i.d. terms $\boldsymbol{\eta}_i + \boldsymbol{\psi}_i$ plus an $o_p\left(n^{-1/2}\right)$ term. We note that the same approach involves nontrivial techniques to apply in high-dimensions. In particular, we discover and exploit the martingale property of the term $\{\omega_i(t) - I(C_i \geq t)\}/G(t)$.

The last piece is the element-wise convergence of the "meat" matrix (19) in the "sandwich" variance estimator. We achieve the following result by repeatedly using Lemmas A.1(ii) and A.2(i).

Lemma 7. *Under conditions (D1)-(D5),* $\sup_{i=1,\ldots,n} \|\widehat{\boldsymbol{\eta}}_i(\widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{\psi}}_i(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\eta}_i - \boldsymbol{\psi}_i\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n} \right) = o_p(1)$. *Hence,* $\|\widehat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}}\|_{\max} = o_p(1)$.

Putting Lemmas 6 and 7 together, we obtain the main result stated in the Theorem below.

Theorem 2. *Let* $\mathbf{c} \in \mathbb{R}^p$ *with* $\|\mathbf{c}\|_1 = 1$ *and* $\mathbf{c}^\top \boldsymbol{\Theta}\boldsymbol{\mathcal{V}}\boldsymbol{\Theta}\mathbf{c} \to \nu^2 \in (0, \infty)$. *Under (D1)-(D5),*

$$\frac{\sqrt{n}\mathbf{c}^\top(\widehat{\mathbf{b}} - \boldsymbol{\beta}^o)}{\sqrt{\mathbf{c}^\top \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\mathcal{V}}}\widehat{\boldsymbol{\Theta}}^\top \mathbf{c}}} \xrightarrow{d} N(0,1).$$

The theorem justifies all the proposed inference and testing procedures in Section 2.4.

4 Simulation Experiments

To assess the finite sample properties of our proposed methods, we conduct extensive simulation experiments with various dimensions and dependence structure among covariates.

4.1 Setup 1

Our first simulation setup follows closely the one of Fine and Gray (1999) but considers high-dimensional covariates. In particular, each $\mathbf{Z}_i$ is a vectors consisting of i.i.d. standard normal random variables. For cause 1, only $\beta_{1,1} = \beta_{1,2} = 0.5$ are non-zero. The cumulative incidence function is:

$$\Pr(T_i \leq t, \epsilon_i = 1 | \mathbf{Z}_i) = 1 - [1 - p\{1 - \exp(-t)\}]^{\exp(\boldsymbol{\beta}_1^\top \mathbf{Z}_i)}.$$

For cause 2, $\beta_{2,1} = \beta_{2,3} = \cdots = \beta_{2,p-1} = -0.5$ and $\beta_{2,2} = \beta_{2,4} = \cdots = \beta_{2,p} = 0.5$, with

$$\Pr(T_i \leq t | \varepsilon_i = 2, \mathbf{Z}_i) = 1 - \exp\left(te^{\boldsymbol{\beta}_2^\top \mathbf{Z}_i}\right).$$

We consider four different combinations: $n = 200$, $p = 300$; $n = 200$, $p = 500$; $n = 200$, $p = 1000$; and $n = 500$, $p = 1000$. Note that this setup considers sparsity for cause 1 but non-sparsity for cause 2 effects. As the Fine-Gray model does not require modeling cause 2 in order to make inference on cause 1, we expect that the non-sparsity in cause 2 effects should not affect the inference on cause 1.

The results are presented in Table 1. We focus on inference for the two non-zero coefficients $\beta_{1,1}$ and $\beta_{1,2}$, as well as one arbitrarily chosen zero coefficient $\beta_{1,10}$. The mean estimates is the average of the one-step $\widehat{\mathbf{b}}$ over the 100 repetitions, reported together with other quantities described below. We can see from the average estimates column that the one-step $\widehat{\mathbf{b}}$ is bias-corrected, and that the presence of many non-zero coefficients for cause 2 does not affect our inference on cause 1.

In practice the choice of the tuning parameters is particularly challenging; the optimal value is determined up to a constant. Moreover, the theoretical results are asymptotic in nature. These together with the finite sample effects of $n \ll p$, lead to suboptimal performance of many proposed one-step correction estimators (van de Geer *et al.*, 2014; Fang *et al.*, 2017). This is more so for survival models, due to the nonlinearity of the loss function and the presence of censoring – both require larger sample size (in order to observe asymptotic statements in the finite samples). In the following we propose a finite-sample correction to the construction of confidence intervals and in particular the estimated standard error (SE).

Let $se(\widehat{b}_j; \widehat{\boldsymbol{\beta}})$ denote the asymptotic standard error as given in section 2.4. As a finite-sample correction we propose to consider $se(\widehat{b}_j; \widehat{\mathbf{b}})$ in place of $se(\widehat{b}_j; \widehat{\boldsymbol{\beta}})$, where the variance estimation based on the initial LASSO estimate $\widehat{\boldsymbol{\beta}}$ is replaced by the one-step $\widehat{\mathbf{b}}$. This can be viewed as another iteration of the bias-correction formula. The resulting SE is therefore a "two-step" SE estimator. We report the coverage rate of the confidence intervals constructed with this finite-sample correction in Table 1 and we observe good coverage close to the nominal level of 95%. We note that with 100 simulation runs the margin of error for the simulated coverage probability is about 2.18%, if the true coverage is 95%; that is, the observed coverage can range between 95+/−4.36%. We note that the coverage is good for all three coefficients, where non-zero or zero. In contrast, results in the existing literature suffer under-coverage of the non-zero coefficients.

The last column 'level/power' in Table 1 refers to the empirical rejection rate of the null hypothesis that the coefficient is zero, by the two-sided Wald test $Z = (\widehat{b}_j - \beta_{1,j})/se(\widehat{b}_j; \widehat{\boldsymbol{\beta}})$ at a nominal 0.05 significance level. We see that although $se(\widehat{b}_j; \widehat{\boldsymbol{\beta}})$ is used, the nominal level is well preserved for the zero coefficient $\beta_{1,10}$, and the power is high for the non-zero coefficients $\beta_{1,1}$ and $\beta_{1,2}$ for the given sample sizes and signal strength.

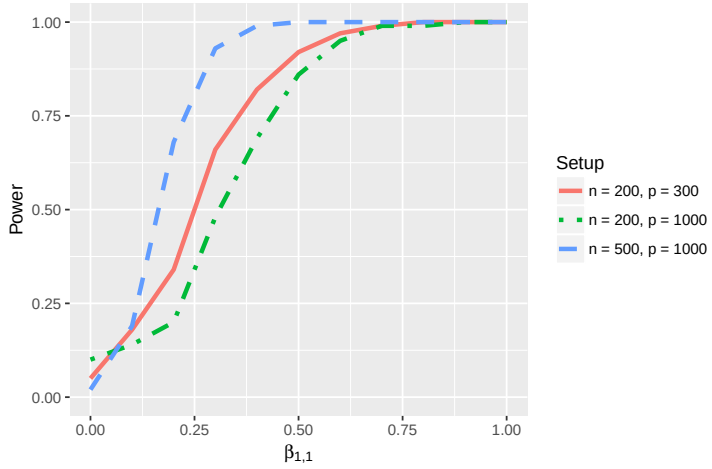
Table 1: Simulation results with independent covariates

	True	Mean Est	SD	SE	SE corrected	Coverage	Level/Power
				n=200, p=300			
$\beta_{1,1}$	0.5	0.51	0.16	0.13	0.25	0.94	0.92
$\beta_{1,2}$	0.5	0.47	0.15	0.14	0.22	0.94	0.93
$\beta_{1,10}$	0	0.03	0.12	0.15	0.18	0.98	0.04

				n=200, p=500			
$\beta_{1,1}$	0.5	0.51	0.16	0.14	0.19	0.93	0.95
$\beta_{1,2}$	0.5	0.48	0.15	0.13	0.19	0.93	0.88
$\beta_{1,10}$	0	-0.01	0.10	0.14	0.16	1.00	0.01
				n=200, p=1000			
$\beta_{1,1}$	0.5	0.46	0.17	0.13	0.18	0.94	0.86
$\beta_{1,2}$	0.5	0.48	0.14	0.13	0.18	0.93	0.92
$\beta_{1,10}$	0	-0.00	0.11	0.14	0.17	0.99	0.06
				n=500, p=1000			
$\beta_{1,1}$	0.5	0.51	0.10	0.08	0.14	0.99	1.00
$\beta_{1,2}$	0.5	0.50	0.10	0.08	0.15	0.99	0.99
$\beta_{1,10}$	0	-0.00	0.07	0.08	0.14	1.00	0.03

We repeat the above simulations with different values for $\beta_{1,1}$ to investigate the power of the Wald test. The results are illustrated in Figure 1, where we see that the power increases with n and decreases with p as expected.

Figure 1: Power curve for testing $\beta_{1,1} = 0$ at nominal level 0.05



4.2 Setup 2

In the second setup we consider the case where the covariates are not all independent, which is more likely the case in practice for high dimensional data. We consider the block dependence structure also used in Binder *et al.* (2009). We consider $n = 500$, $p = 1000$; $\beta_{1,1\sim8} = 0.5$, $\beta_{1,9\sim12} = -0.5$ and the rest are all zero. $\beta_{2,1\sim4} = \beta_{2,13\sim16} = 0.5$, $\beta_{2,5\sim8} = -0.5$ and the rest of $\boldsymbol{\beta}_1$ are all zero. The covariates are grouped into four blocks of size 4, 4, 8 plus the rest, with the within-block correlations equal to 0.5, 0.35, 0.05 and 0. The four blocks are separated by the horizontal lines in Table 2.

Table 2 shows the inferential results for the non-zero coefficients $\beta_{1,1} \sim \beta_{1,12}$, as well as the zero coefficients $\beta_{1,13} \sim \beta_{1,16}$ from the third correlated block that also contains some of the non-zero coefficients, and plus arbitrarily chosen zero coefficient $\beta_{1,30}$. The initial LASSO estimator tended to select only one of every four non-zero coefficients of the correlated covariates (data not shown), as it is known that block dependence structure is particularly challenging for the Lasso type estimators. On the other hand, the one-step estimator performed remarkably well, capturing all of the non-zero coefficients.

Compared to the results in the last part of Table 1 with the same n and p, the block correlated covariates led to slightly more bias in $\widehat{\mathbf{b}}$, although the CI coverage remained high. The power also remained high, although in the third block with the mixed signal and noise variables the type I error rates appeared slightly high.

Table 2: Simulation results with block correlated covariates

	True	Mean Est	SD	SE	SE corrected	Coverage	Level/Power
				n=500, p=1000			
$\beta_{1,1}$	0.5	0.47	0.10	0.07	0.12	0.97	1.00
$\beta_{1,2}$	0.5	0.48	0.10	0.07	0.12	0.94	0.98
$\beta_{1,3}$	0.5	0.47	0.10	0.07	0.12	0.98	1.00
$\beta_{1,4}$	0.5	0.47	0.10	0.07	0.12	0.94	1.00
$\beta_{1,5}$	0.5	0.48	0.10	0.06	0.11	0.93	1.00
$\beta_{1,6}$	0.5	0.46	0.10	0.06	0.11	0.94	1.00
$\beta_{1,7}$	0.5	0.47	0.09	0.06	0.11	0.95	1.00
$\beta_{1,8}$	0.5	0.47	0.08	0.06	0.11	0.98	1.00
$\beta_{1,9}$	-0.5	-0.44	0.08	0.06	0.11	0.93	1.00
$\beta_{1,10}$	-0.5	-0.42	0.08	0.06	0.11	0.92	1.00
$\beta_{1,11}$	-0.5	-0.41	0.08	0.06	0.11	0.91	1.00
$\beta_{1,12}$	-0.5	-0.43	0.07	0.05	0.11	0.94	1.00
$\beta_{1,13}$	0	-0.01	0.06	0.05	0.11	0.98	0.11
$\beta_{1,14}$	0	-0.02	0.05	0.05	0.11	1.00	0.06
$\beta_{1,15}$	0	-0.02	0.06	0.06	0.11	0.99	0.08
$\beta_{1,16}$	0	-0.02	0.06	0.05	0.11	1.00	0.05
$\beta_{1,30}$	0	-0.00	0.05	0.06	0.11	1.00	0.01

5 SEER-Medicare data example

The SEER-Medicare linked database contains clinical information and claims codes for 57011 patients diagnosed between 2004 and 2009. The clinical and demographic information were collected at diagnosis, and the insurance claim data were from the year prior to diagnosis. The clinical information contained PSA, Gleason Score, AJCC stage and year of diagnosis. Demographic information included age, race, and marital status. The same data set was considered in Hou *et al.* (2017), where the emphasis was on variable selection and prediction error. Our focus is on testing and construction of confidence intervals.

There were 9326 binary claims codes in the data. Here we would like to identify the risk factors for non-cancer mortality using the Fine-Gray model. We kept only the claims codes with at least 10 and at most 1990 occurrences. The resulting dataset had 1197 covariates. In the following we consider 413 patients diagnosed during the year of 2004. Among them 76 died from the cancer and 337 had deaths unrelated to cancer. We center and standardize all the covariates before performing the analysis. To determine the penalty parameters λ and λ_j we used 10-fold cross-validation.

In Table 3 we present the result for a selected set of the coefficients (due to space the set of all variables is not presented in full although we have computed 1197 p-values). We presented some variables with the larges p-values and some that Lasso (initial estimator) returned as zero initially. We also focused on heart disease and colon cancer as potential significan non-cancer mortality (different than prostate cancer) as well as prostate cancer variables. A descriptions of the variables is given in Table 4. For each coefficient, we report the initial LASSO estimate $\widehat{\boldsymbol{\beta}}$, one-step estimate $\widehat{\mathbf{b}}$, corrected SE, the 95% CI constructed with the corrected SE and the Wald test p-value (2-sided) calculated using the uncorrected SE.

Table 3: Inference for the SEER-Medicare linked data on non-cancer mortality among prostate cancer patients

Variables	Initial estimate	One-step estimate and Inference			
	$\widehat{\beta}$	$\widehat{b}$	se$(\widehat{b})$	95% CI	p-value
Age	0.075	0.096	0.009	[0.078, 0.114]	2e-24*
Marital	0	0.218	0.147	[-0.071, 0.507]	0.042
Race.OvW	0	-0.213	0.224	[-0.652, 0.225]	0.317
Race.BvW	0.244	0.528	0.122	[0.288, 0.767]	1e-04*
PSA	0	0.005	0.003	[-0.000, 0.010]	0.041
GleasonScore	0	0.084	0.050	[-0.014, 0.182]	0.085
AJCC-T2	0	-0.130	0.146	[-0.418, 0.157]	0.218
ICD-9 51881	0.866	1.357	0.361	[0.650, 2.064]	4e-07*
ICD-9 4280	0.404	0.697	0.062	[0.576, 0.818]	2e-06*
CPT 93015	-0.061	-1.042	0.327	[-1.683, -0.401]	4e-05*
ICD-9 42731	0.135	0.459	0.191	[0.086, 0.833]	0.001*
CPT 72050	0	3.718	0.208	[3.310, 4.125]	4e-05*
ICD-9 6001	0	-2.454	0.577	[-3.585, -1.322]	0.000*
CPT 74170	0	-1.689	0.288	[-2.255, -1.124]	0.001*
ICD-9 2948	0.539	0.746	0.205	[0.343, 1.148]	0.009
ICD-9 49121	0.150	0.476	0.215	[0.055, 0.896]	0.015
ICD-9 2989	0.079	0.450	0.135	[0.184, 0.715]	0.062
ICD-9 79093	-0.056	-0.348	0.176	[-0.693, -0.002]	0.088
ICD-9 41189	0	1.332	0.434	[0.480, 2.184]	0.003**
CPT 45380	0	-2.250	0.544	[-3.318, -1.182]	0.003**
ICD-9 3320	0	0.378	0.373	[-0.353, 1.110]	0.327

$*$ denotes 5% significance after Bonferoni correction
$**$ denotes 10% significance after Bonferoni correction

Table 4: Description of the variables in Table 3

Code	Description
Age	Age at diagnosis
Marital	marSt1: married vs other
Race.OvW	Race: Other vs White
Race.BvW	Race: Black with White
PSA	PSA
GleasonScore	Gleason Score
AJCC-T2	AJCC stage-T: T2 vs T1
ICD-9 51881	Acute respiratry failure (Acute respiratory failure)
ICD-9 4280	Congestive heart failure; nonhypertensive [108.]
CPT 93015	Global Cardiovascular Stress Test
ICD-9 42731	Cardiac dysrhythmias [106.]
CPT 72050	Diagnostic Radiology (Diagnostic Imaging) Procedures of the Spine and Pelvis

ICD-9 6001	Nodular prostate
CPT 74170	Diagnostic Radiology (Diagnostic Imaging) Procedures of the Abdomen
ICD-9 2948	Delirium dementia and amnestic and other cognitive disorders [653]
ICD-9 49121	Obstructive chronic bronchitis
ICD-9 2989	Unspecified psychosis
ICD-9 41189	acute and subacute forms of ischemic heart disease, other
CPT 45380	Under Endoscopy Procedures on the Rectum
ICD-9 3320	Parkinsons disease [79.]

In Table 3 we see that the claims codes ICD-9 4280, CPT 93015, ICD-9 42731 are all related to the heart disease, and are all significant at 5% level Bonferonni correction for the 21 variables included in the table. A heart attack indicator variable, ICD-9 41189, shows up significant at 10% level (again with correction for multiple testing). An indicator of a disease in the abdomen, CPT 74170, is significant at 5% although the initial Lasso regularized method failed to include such variable. Similar result is seen for the indicator of a fall (CPT 72050) which for an elderly person can be fatal. An indicator of a colon cancer (CPT 45380) turns out to be significant at 10% although the Lasso method set it to zero initially. Therefore, our one-step method is able to recover important risk factors that would be missed by the initial LASSO estimator.

In contrast, non-life-threatening diseases, as expected, are not significant for non-cancer mortality. These include Parkinson's (ICD-9 3320), Psychosis (ICD-9 2989), Bronchitis (ICD-9 49121) and Dementia (ICD-9 2948) in the table. We also note that the prostate cancer related variables, PSA, Gleason Score, AJCC and ICD-9 6001, all have large p-values for non-cancer mortality. This is consistent with the results in Hou *et al.* (2017), where under the competing risk models the predictors for a second cause is only secondary in importance in predicting the events due to the first cause.

6 Discussion

This article focuses on estimation and inference under the Fine-Gray model with many more covariates than the number of events, which is well-known to be the effective sample size for survival data. The article proposes conditions under which a Lasso estimator performs well in terms of prediction error. It also develops a new one-step estimator that can be utilized for asymptotically optimal inference: confidence intervals and testing.

An often overlooked restriction on the time-dependent covariates $Z_i(t)$, $i = 1, \ldots, n$, under the Fine-Gray model is that $Z_i(t)$ must be observable even after the i-th subject experiences a type 2 event. In practice, $Z_i(t)$ should be either time independent or external (Kalbfleisch and Prentice, 2002). In our case the continuity conditions (C2) and (D3) are easily satisfied if the majority of the elements in $Z_i(t)$ are time independent, which is most likely to be the case in practice. Our theory does not apply in studies involving longitudinal variables that are supposed to be truly measured continuously over time.

We have illustrated that the method based on regularization only (without bias correction) might have serious disadvantages in many complex data situations – for example, it may potentially fail to identify important variables that are associated with the response. From the analysis of the SEER-medicare data we see that variables like CPT 72050 (related to fall) or CPT 74170 (related to diagnostic imaging of the abdomen, often for suspected malignancies) would not have been discovered as important risk factors for non-cancer mortality by regularization alone. In reality,

both can be life threatening events for an elderly patient. The one-step estimate, on the other hand, was able to detect these therefore providing a valuable tool for practical applications. The one-step estimator is applicable as long as the model is sparse, and no minimum signal strength is required; this is another important aspect which makes the estimator more desirable for practical use than the LASSO type estimators.

Acknowledgement

We would like to acknowledge our collaboration with Dr. James Murphy of the UC San Diego Department of Radiation Medicine and Applied Sciences on the linked Medicare-SEER data analysis project that motivated this work. We would also like to thank his group for help in preparing the data set.

Supplementary Materials

In this section we provide details of all of the theoretical results. Preseding the proof of each statement we present the needed preliminary lemmas (numbered in letters).

A Concentration Inequalities for Time-dependent Processes

Lemma A.1. *Let* $\{(\mathbf{S}_i(t), N_i(t)) \in \mathbb{R}^q \times \mathbb{N} : i = 1, \ldots, n, t \in [0, t^*]\}$ *be i.i.d. pairs of random processes. Each* $N_i(t)$ *is a counting process bounded by* K_N*. Denote its jumps as* $0 \leq t_{i1} < \cdots < t_{iK_i} \leq t^*$*. Let* $\bar{\mathbf{S}}(t) = n^{-1}\sum_{i=1}^n \mathbf{S}_i(t)$ *and* $\mathrm{E}\{\mathbf{S}_i(t)\} = \mathbf{s}(t)$*. Suppose* $\sup_{1\leq i<j\leq n}\sup_{t\in[0,t^*]} \|\mathbf{S}_i(t) - \mathbf{S}_j(t)\|_{\max} \leq K_S$ *almost surely. Then,*

(i) $\Pr\left(\sup_{i=1,\ldots,n}\sup_{j=1,\ldots,K_i} \left\|\bar{\mathbf{S}}(t_{ij}) - \mathbf{s}(t_{ij})\right\|_{\max} > K_S x + (K_S)/n\right) < 2nK_N q e^{-nx^2/2}$.

(ii) Assume in addition that each $\mathbf{S}_i(t)$ *is càglàd generated by*

$$\mathbf{S}_i(t) = \mathbf{S}_i(0) + \int_0^t \mathbf{d}_s(u)du + \int_0^t \mathbf{J}_s(u)dN_i(u)$$

for some $\mathbf{d}_s(t)$ *and* $\mathbf{J}_s(t)$ *satisfying* $\|\mathbf{d}_s(t)\|_{\max} < L_S$ *and* $\|\mathbf{J}_s(u)\|_{\max} < K_S$*, and* $\mathrm{E}\{N_i(t)\} = \int_0^t h_i^N(u)du$ *for some* $h_i^N(t) \leq L_N$*. We have*

$$\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]} \left\|\bar{\mathbf{S}}(t) - \mathbf{s}(t)\right\|_{\max} = O_p(\sqrt{\log(nK_N q)/n}).$$

Lemma A.2. *Let* $\{M_i(t) : t \in [0, t^*],\ i = 1, \ldots, n\}$ *be a* $\mathcal{F}_t$*-adapted counting process martingales* $M_i(t) = N_i(t) - \int_0^t Y_i(t)h_i(u)du$ *satisfying* $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]} h_i(t) \leq K_h$*. Let* $\{\mathbf{\Phi}_i(t) : t \in [0, t^*],\ i = 1, \ldots, n\}$ *be the* q *dimensional* $\mathcal{F}_{t-}$*-measurable processes such that*

$$\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]} \|\mathbf{\Phi}_i(t)\|_{\max} \leq K_\Phi.$$

For $\mathbf{M}_\Phi(t) = n^{-1}\sum_{i=1}^n \int_0^t \mathbf{\Phi}_i(u)dM_i(u)$ *, we have*

(i) $\Pr\left(\sup_{t\in[0,t^*]} \|\mathbf{M}_\Phi(t)\|_{\max} \geq K_\Phi(1 + K_h t^*)x + K_\Phi K_h t^*/n\right) \leq 2qe^{-nx^2/4}$.

(ii) Assume in addition $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]} \|\mathbf{\Phi}_i(t)\|_{\max} = O_p(a_n)$ *and* $K_h t^* \asymp O(1)$*. Then,* $\sup_{t\in[0,t^*]} \|\mathbf{M}_\Phi(t)\|_{\max} = O_p(a_n\sqrt{\log(q)/n})$.

B Proofs of Main Results

We state the related preliminary results before the proof of each of the main results. The proofs of the preliminary results are given in the next section.

Proof of Lemma A.1.

(i) Without loss of generality, let t_{11} be the first jump time of $N_1(t)$. By the i.i.d. assumption, t_{11} is independent of all $\mathbf{S}_i(t)$ with $i \geq 2$. Thus, the sequence

$$\mathbf{L}_l = n^{-1} \sum_{i=2}^{l} \{\mathbf{S}_i(t_{11}) - \mathbf{s}(t_{11})\}$$

is a martingale with respect to filtration $\big\{\sigma\big(\mathbf{S}_i(t), i \leq l\big), l = 2, \ldots, n\big\}$. The increment is bounded as

$$n^{-1} \{\mathbf{S}_i(t_{11}) - \mathbf{s}(t_{11})\} = n^{-1} \mathrm{E}_{\mathbf{S}_j} \{\mathbf{S}_i(t_{11}) - \mathbf{S}_j(t_{11})\} \leq n^{-1} K_S.$$

Applying Azuma (1967) inequality to $\mathbf{L}_n$, we get $\Pr\left(\|\mathbf{L}_n\|_{\max} > K_S x\right) < 2qe^{-nx^2/2}$. Since the dropped first term is also bounded by K_S/n, we get

$$\Pr\left(\left\|\bar{\mathbf{S}}(t_{11}) - \mathbf{s}(t_{11})\right\|_{\max} > K_S x + K_S/n\right) < 2qe^{-nx^2/2}.$$

We use simple union bound to extend the result to all t_{ij}'s whose number is at most nK_N.

(ii) Define a deterministic set $\mathcal{T}_n = \{kt^*/n : k = 1, \ldots, n\} \cup \mathcal{T}_z$. By the union bound of Hoeffding (1963) inequality, we have

$$\Pr\left(\sup_{t \in \mathcal{T}_n} \left\|\bar{\mathbf{S}}(t) - \mathbf{s}(t)\right\|_{\max} > K_S x\right) < 2(n + |\mathcal{T}_z|)qe^{-nx^2/2}.$$

Combining the result from Lemma A.1(i), we obtain

$$\left\|\bar{\mathbf{S}}(t) - \mathbf{s}(t)\right\|_{\max} = O_p(\sqrt{\log(npq)/n})$$

over a grid containing $\mathcal{T}_n$ and jumps of $N_i(t)$. We only need to show that the variation of $\bar{\mathbf{S}}(t) - \mathbf{s}(t)$ is sufficiently small inside each bin created by the grid.

Let t' and t'' be consecutive elements by order in $\mathcal{T}_n$. By our construction, there is no jump of any of the counting processes $N_i(t)$ in the interval (t', t''). Otherwise, the jump time is another element in $\mathcal{T}_n$ between t' and t'' so that t' and t'' are not consecutive elements by order. Under the assumption of the lemma, elements of all $\mathbf{S}_i(t)$'s are L_S-Lipschitz in (t', t''). Moreover, $|t'' - t'| \leq t^*/n$ because of the deterministic $\{kt^*/n : k = 1, \ldots, n\} \subset \mathcal{T}_n$. Along with the càglàd property, we obtain a bound of variation of $\bar{\mathbf{S}}(t)$ in (t', t'')

$$\sup_{t \in (t', t'')} \|\bar{\mathbf{S}}(t) - \bar{\mathbf{S}}(t'')\|_{\max} \leq \sup_{i=1,\ldots,n} \sup_{t \in (t', t'')} \|\mathbf{S}_i(t) - \mathbf{S}_i(t'')\|_{\max} \leq L_S |t'' - t'| \leq L_S t^*/n.$$

For any $t \in (t', t'')$, we bound the variation of $\mathbf{s}(t)$ by

$$\|\mathbf{s}(t) - \mathbf{s}(t'')\|_{\max} \leq \int_t^{t''} \mathrm{E} \|\mathbf{d}_s(u)\|_{\max} du + \int_t^{t''} \mathrm{E}\{\|\mathbf{J}_s(u)\|_{\max} h_i^N(u)\} du \leq (L_S + K_S L_N) t^*/n.$$

For arbitrary $t \in [0, t^*]$, we find the the corresponding bin $(t', t'']$ contains t. Putting the results together, we have

$$\|\bar{\mathbf{S}}(t)-\mathbf{s}(t)\|_{\max} \leq \|\bar{\mathbf{S}}(t)-\bar{\mathbf{S}}(t'')\|_{\max}+\|\mathbf{s}(t)-\mathbf{s}(t'')\|_{\max}+\|\bar{\mathbf{S}}(t'')-\mathbf{s}(t'')\|_{\max} \leq O_p(\sqrt{\log(npq)/n})+O(1/n).$$

□

Proof of Lemma A.2.

(i) The summands in $\mathbf{M}_\Phi(t)$ are the integrals of $\mathcal{F}_{t-}$-measurable processes over $\mathcal{F}_t$-adapted martingales, so $\mathbf{M}_\Phi(t)$ is a $\mathcal{F}_t$-adapted martingale (Kalbfleisch and Prentice, 2002).

Suppose $\{T_i : i = 1, \ldots, n\}$ are the jump times of $\{N_i(t)\}$. We artificially set $T_i = t^*$ if $N_i(t)$ has no jump in $[0, t^*]$. Define $0 \leq R_1 \leq \cdots \leq R_{K_N+n}$ be the order statistics of $\{T_i : i = 1, \ldots, n\} \cup \{kt^*/n : k = 1, \ldots, n\}$. Hence, $\{R_k : k = 1, \ldots, 2n\}$ is a set of ordered $\mathcal{F}_t$ stopping times. Applying optional stopping theorem, we get a discrete time martingale $\mathbf{M}_\Phi(R_k)$ adapted to $\mathcal{F}_{R_k}$.

The increment of $\mathbf{M}_\Phi(R_k)$ comes from either the counting part or the compensator part, which we can bound separately. By our construction of R_k's, each left-open right-closed bin $(R_k, R_{k+1}]$ satisfies two conditions. There is at most one jump from $\sum_{i=1}^n N_i(t)$ in the bin at R_{k+1}. The length of the bin is at most t^*/n. The increment of the martingale $\mathbf{M}_\Phi(t)$ over $(R_k, R_{k+1}]$ is decomposed into two coordinate-wise integrals, a jump minus a compensator,

$$\mathbf{M}_\Phi(t) = n^{-1} \sum_{i=1}^n \int_{R_k}^{R_{k+1}} \mathbf{\Phi}_i(u) dN_i(u) - n^{-1} \sum_{i=1}^n \int_{R_k}^{R_{k+1}} \mathbf{\Phi}_i(u) h_i(u) du.$$

With the assumed a.s. upper bound for $\sup_{t\in[0,t^*]} \|\mathbf{\Phi}_i(t)\|_{\max} \leq K_\Phi$, we have almost surely the jump of $\mathbf{M}_\Phi(t)$ in the bin be bounded by

$$\left\| n^{-1} \sum_{i=1}^n \int_{R_k}^{R_{k+1}} \mathbf{\Phi}_i(u) dN_i(u) \right\|_{\max} \leq K_\Phi/n.$$

Additionally with the assumed upper bound for $\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} h_i(t) \leq K_h$, we have the compensator of $\mathbf{M}_\Phi(t)$ increases over the bin by at most

$$\left\| \int_{R_k}^{R_{k+1}} n^{-1} \sum_{i=1}^n \mathbf{\Phi}_i(u) h_i(u) du \right\|_{\max} \leq K_\Phi K_h (R_{k+1} - R_k) \leq K_\Phi K_h t^*/n.$$

We obtain a uniform concentration inequality for $\mathbf{M}_\Phi(R_k)$ by Azuma (1967)

$$\Pr\left(\sup_{k=1,\ldots,2n} \|\mathbf{M}_\Phi(R_k)\|_{\max} \geq K_\Phi(1 + K_h t^*)x \right) \leq 2qe^{-nx^2/4}.$$

Remark that the uniform version of Azuma (1967) is the application of Doob's maximal inequality (Durrett, 2013, Theorem 5.4.2, page 213). For $t \in (R_k, R_{k+1})$, we use the bounded increment derived above

$$\|\mathbf{M}_\Phi(t) - \mathbf{M}_\Phi(R_k)\|_{\max} \leq \left\| \int_{R_k+}^{R_{k+1}+} n^{-1} \sum_{i=1}^n \mathbf{\Phi}_i(u) h_i(u) du \right\|_{\max} \leq K_\Phi K_h t^*/n.$$

(ii) Under the additional assumption $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\|\boldsymbol{\Phi}_i(t)\|_{\max}=O_p(a_n)$, we can find $K_{\Phi,\varepsilon}$ for every $\varepsilon>0$ such that

$$\Pr\left(\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\|\boldsymbol{\Phi}_i(t)\|_{\max}\le K_{\Phi,\varepsilon}a_n\right)\ge 1-\varepsilon/2$$

for any n. We apply Lemma A.2(i) to obtain that event

$$\left\{\sup_{t\in[0,t^*]}\|\mathbf{M}_\Phi(t)\|_{\max}\le K_{\Phi,\varepsilon}a_n\{(1+K_ht^*)\sqrt{2\log(4q)/n}+K_ht^*/n\},\ \sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\|\boldsymbol{\Phi}_i(t)\|_{\max}\le K_{\Phi,\varepsilon}a_n\right\}$$

occurs with probability no less than $1-\varepsilon$.

□

Lemma B.4*. *Under (D2),* $\sup_{t\in[0,t^*]}|n/\{\sum_{i=1}^n I(X_i\ge t^*)\}|$, $\sup_{t\in[0,t^*]}|S^{(0)}(t,\boldsymbol{\beta}^o)^{-1}|$ *and* $\sup_{t\in[0,t^*]}|\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)^{-1}|$ *are all* $O_p(1)$.

Lemma B.3. *Let* $\{a_i(t):t\in[0,t^*],i=1,\ldots,n\}$ *be a set of nonnegative processes. Under* (39),

$$\left\|\frac{\sum_{i=1}^n a_i(t)\mathbf{Z}_i(t)^{\otimes l}}{\sum_{i=1}^n a_i(t)}\right\|_{\max}\le(K/2)^l,\ \text{and}\ \left\|\frac{\mathrm{E}\{a_i(t)\mathbf{Z}_i(t)^{\otimes l}\}}{\mathrm{E}\{a_i(t)\}}\right\|_{\max}\le(K/2)^l.$$

As a result, the maximal norms of $\mathbf{S}^{(l)}(t,\boldsymbol{\beta})/S^{(0)}(t,\boldsymbol{\beta})$ *and* $\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta})/\widetilde{S}^{(0)}(t,\boldsymbol{\beta})$ *and* $\mathbf{s}^{(l)}(t,\boldsymbol{\beta})/s^{(0)}(t,\boldsymbol{\beta})$ *are all uniformly bounded by* $(K/2)^l$.

Lemma B.4. *Define* $\widetilde{S}^{(0)}(t;M)=n^{-1}\sum_{i=1}^n I(C_i\ge t^*)Y_i(t^*)\min\{M,e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\}$. *Let* $T^1_{(1)},\ldots,T^1_{(K_N)}$ *be the observed type-1 events. Under (C1), the event*

$$\Omega_{r_*}=\left\{n^{-1}\sum_{i=1}I(X_i\ge t^*)\ge r_*/(2M),\ \sup_{k\in1\ldots K_N}\widetilde{S}^{(0)}\left(T^1_{(k)};M\right)\ge r_*/2\right\}\tag{B.1}$$

occurs with probability at least $1-e^{-nr_*^2/(2M^2)}-ne^{-n(r_*-2/n)^2/(8M^2)}$.

On Ω_{r_*}, *we have* $\sup_{k\in1\ldots K_N}\widetilde{S}^{(0)}(T^1_{(k)})\ge r_*/2$.

Lemma B.5. *Define*

$$K_{e,\varepsilon}=\frac{e^{c_zL_z\|\boldsymbol{\beta}^o\|_\infty D}}{Dh_*}\log(n/\varepsilon)\tag{B.2}$$

Under (C2), the event

$$\Omega_{e,\varepsilon}=\left\{\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}I(\delta_i\epsilon_i>1)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}<K_{e,\varepsilon}\right\}\tag{B.3}$$

occurs with probability at least $1-\varepsilon$.

Lemma B.6. *Define the IPW weights with true* $G(t)$ *as* $\widetilde{\omega}_i(t)=r_i(t)G(t)/G(X_i\wedge t)$ *and*

$$C^\omega_{n,p,\varepsilon}=4(M/r_*)^2\left\{(1+K_ct^*)\sqrt{4\log(2/\varepsilon)/n}+K_ct^*/n\right\}.$$

On event Ω_{r_*}, $\Omega_{KM,\varepsilon}=\left\{\sup_{t\in[0,t^*]}\sup_{t\in[0,t^*]}|\omega_i(t)-\widetilde{\omega}_i(t)|\le C^\omega_{n,p,\varepsilon}\right\}$ *occurs with probability at least* $\Pr(\Omega_{r_*})-\varepsilon$.

Lemma B.7. *Define* $\boldsymbol{\Delta}^{(l)}(t) = \mathbf{S}^{(l)}(t, \boldsymbol{\beta}^o) - \widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta}^o)$. *Let* $T^1_{(1)}, \ldots, T^1_{(K_N)}$ *be the observed type-1 events for some* $K_N \leq n$. *Denote*

$$C^{(l)}_{n,p,\varepsilon} = \frac{K_{e,\varepsilon} K^l}{2^l} \left\{ \frac{4M^2(1+K_c t^*)}{r_*^2} \sqrt{\frac{4\log(2/\varepsilon)}{n}} + \frac{4M^2 K_c t^*}{r_*^2 n} + \sqrt{\frac{2\log(2np^l/\varepsilon)}{n}} + \frac{1}{n} \right\}$$

as in (45) *Under (C1) and (C2),*

$$\Omega_{\Delta,\varepsilon} = \left\{ \max_{l=0,1,2} \sup_{k \in 1 \ldots K_N} \left\| \boldsymbol{\Delta}^{(l)} \left(T^1_{(k)} \right) \right\|_{\max} \leq C^{(l)}_{n,p,\varepsilon} \right\} \cap \Omega_{r_*} \cap \Omega_{e,\varepsilon} \cap \Omega_{KM,\varepsilon}, \tag{B.4}$$

with Ω_{r_*}, $\Omega_{e,\varepsilon}$ *and* $\Omega_{KM,\varepsilon}$ *defined in Lemmas B.4, B.5 and B.6, occurs with probability at least* $1 - e^{-nr_*^2/(2M^2)} - ne^{-n(r_*-2/n)^2/(8M^2)} - 5\varepsilon$.

On $\Omega_{\Delta,\varepsilon}$, *we have for* $l = 1, 2$,

$$\sup_{k \in 1 \ldots K_N} \left\| \frac{\mathbf{S}^{(l)} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right)}{S^{(0)} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right)} - \frac{\widetilde{\mathbf{S}}^{(l)} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right)}{\widetilde{S}^{(0)} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right)} \right\|_{\max} \leq 2\{ C^{(l)}_{n,p,\varepsilon} + (K/2)^l C^{(0)}_{n,p,\varepsilon} \} / r_*.$$

Proof of Lemma 1.

Let $T^1_{(1)}, \ldots, T^1_{(K_N)}$ be the observed type-1 events. We may decompose the score $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ as its martingale proxy plus an approximation error,

$$\dot{\mathbf{m}}(\boldsymbol{\beta}^o) = \dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o) + n^{-1} \sum_{k=1,\ldots,K_N} \left\{ \widetilde{\mathbf{Z}} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right) - \bar{\mathbf{Z}} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right) \right\}.$$

Recall that the counting process for observed type-1 event can be written as $N^o_i(t) = \int_0^t I(C_i \geq u) dN^1_i(t)$. Moreover, $\dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$ takes the form of the Cox model score with counting process $\{N^o_i(t)\}$ and at-risk process $\{I(C_i \geq t) Y_i(t)\}$. The "censoring complete" filtration $\mathcal{F}^*_t$ can also be equivalently generated by $\{N^o_i(t), I(C_i \geq t) Y_i(t), \mathbf{Z}_i(t)\}$. Thus, we may apply Huang *et al.* (2013) Lemma 3.3 under (39) from (C1),

$$\Pr(\|\dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)\|_\infty > Kx) \leq 2pe^{-nx^2/2}.$$

Notice that the inequality is sharper than that in Lemma A.2(i) because the compensator part of $\dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$ is zero.

The concentration result for approximation error

$$\widetilde{\mathbf{Z}} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right) - \bar{\mathbf{Z}} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right) = \frac{\mathbf{S}^{(1)} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right)}{S^{(0)} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right)} - \frac{\widetilde{\mathbf{S}}^{(1)} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right)}{\widetilde{S}^{(0)} \left(T^1_{(k)}, \boldsymbol{\beta}^o \right)}$$

is established in Lemma B.7 on $\Omega_{\Delta,\varepsilon}$. We obtain the concentration inequality for $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ by adding the bounds and tail probabilities together. □

Proof of Lemma 2.

Our strategy here is the same as that for Lemma 1. We first show that $\kappa(\xi, \mathcal{O}; \ddot{\mathbf{m}}(\boldsymbol{\beta}^o))$ is lower bounded by $\kappa(\xi, \mathcal{O}; -\ddot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o))$ plus a diminishing error. Since $\ddot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$ takes the form of a Cox model Hessian, we then may apply the results from Huang *et al.* (2013).

By Lemma 4.1 in Huang *et al.* (2013) (for a similar result, see van de Geer and Bühlmann (2009) Corollary 10.1),

$$\kappa^2(\xi, \mathcal{O}; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)) \geq \kappa^2(\xi, \mathcal{O}; -\ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)) - s_o(\xi+1)^2 \|\ddot{\mathbf{m}}(\boldsymbol{\beta}^o) - \ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)\|_{\max}.$$

Let $T^1_{(1)}, \ldots, T^1_{(K_N)}$ be the observed type-1 events. We can write $\ddot{\mathbf{m}}(\boldsymbol{\beta}^o) - \ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$ as

$$-n^{-1} \sum_{k=1}^{K_N} \left[\frac{\mathbf{S}^{(2)}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)}{S^{(0)}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)} - \frac{\widetilde{\mathbf{S}}^{(2)}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)}{\widetilde{S}^{(0)}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)} - \bar{\mathbf{Z}}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)^{\otimes 2} + \widetilde{\mathbf{Z}}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)^{\otimes 2} \right].$$

By Lemma B.3, $\sup_{t\in[0,t^*]} \|\bar{\mathbf{Z}}(t, \boldsymbol{\beta}^o)\|_\infty$ and $\sup_{t\in[0,t^*]} \|\widetilde{\mathbf{Z}}(t, \boldsymbol{\beta}^o)\|_\infty$ are both bounded by $K/2$. On the $\Omega_{\Delta,\varepsilon}$, we apply Lemma B.7 once with $l = 2$ and twice with $l = 1$ to get $\|\ddot{\mathbf{m}}(\boldsymbol{\beta}^o) - \ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)\|_{\max} \leq \left\{ 2C^{(2)}_{n,p,\varepsilon} + 4KC^{(1)}_{n,p,\varepsilon} + (5/2)K^2 C^{(0)}_{n,p,\varepsilon} \right\} / r_*$.

Our (C1) and (C3) contains all the condition for Theorem 4.1 in Huang *et al.* (2013). Hence, we may apply their result

$$\kappa^2(\xi, \mathcal{O}; -\ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)) \geq \kappa^2(\xi, \mathcal{O}; \boldsymbol{\Sigma}(M)) - s_o(\xi+1)^2 K^2 \left\{ (1 + t^* K_h) \sqrt{2 \log\left(p(p+1)/\varepsilon\right)/n} + (2/r_*) t^* K_h t^2_{n,p,\varepsilon} \right\}$$

with probability at least $\Pr(\Omega_{\Delta,\varepsilon}) - 3\varepsilon$. We have bounded $\widetilde{S}^{(0)}(t; M)$ away from zero at all observed type-1 events in $\Omega_{\Delta,\varepsilon}$, so the $e^{-nr_*^2/(8M^2)}$ term is absorbed into $\Pr(\Omega_{\Delta,\varepsilon})$. □

Proof of Theorem 1.

Observe that the same techniques as those of Huang *et al.* (2013) apply (see for example Lemmas 3.1 and 3.2 therein). The structure of the partial likelihood is the same as that of the Cox model modular the IPW weight functions $w_j(t)$. Following the same line of proof we can easily obtain

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \leq \frac{e^{\varsigma}(\xi+1) s_o \lambda}{2\kappa\big(\xi, \mathcal{O}; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big)^2} \tag{B.5}$$

with $\varsigma_{\mathbf{b}} = \sup_{t\in[0,t^*]} \sup_{1\leq i<j\leq n} |\mathbf{b}^\top \{\mathbf{Z}_i(t) - \mathbf{Z}_j(t)\}|$ in the event $\|\dot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_1 \leq \lambda(\xi-1)/(\xi+1)$. The proof is then completed by utilizing Lemma 1 and Lemma 2.

□

Proof of Theorem 1*.

Since we assume (D1) now, the relative risks are bounded almost surely from above and below by constants $0 < e^{-K_b} \leq e^{\boldsymbol{\beta}^{o\top} \mathbf{Z}_i(t)} \leq e^{K_b} < \infty$. We may set $M = e^{K_b}$ to directly obtain (C1) and (C3) from (D2) and (D4). We can also improve the rate of estimation error in Theorem 1 by $\log(n)$ because we need not let $K_{e,\varepsilon}$ in Lemma B.7 to grow with n. □

Proof of Lemma 3.

Denote

$$\boldsymbol{U} = \int_0^{t^*} \{\mathbf{Z}(t) - \boldsymbol{\mu}(t)\} dN^o(t).$$

Without loss of generality, we set $j = 1$. Since we define $\boldsymbol{\gamma}_1^* = \operatorname{argmin}_{\boldsymbol{\gamma}} \bar{\Gamma}(\boldsymbol{\gamma})$ as the minimizer of a convex function, it must satisfy the first order condition

$$\nabla_\gamma \bar{\boldsymbol{\Gamma}}(\boldsymbol{\gamma}_1^*) = \mathrm{E}\left\{ (U_1 - \boldsymbol{U}_{-1}^\top \boldsymbol{\gamma}_1^*) \boldsymbol{U}_{-1} \right\} = \mathbf{0}_{p-1}.$$

Recall that $\tau_1^2 = \bar{\Gamma}(\boldsymbol{\gamma}_1^*)$. Applying the first order condition, we get

$$\tau_1^2 = \mathrm{E}\{U_1 - \boldsymbol{U}_{-1}^\top \boldsymbol{\gamma}_1^*\}^2 = \mathrm{E}\{(U_1 - \boldsymbol{U}_{-1}^\top \boldsymbol{\gamma}_1^*)U_1\}.$$

We construct a vector $\boldsymbol{\theta}_1 = (1, -\boldsymbol{\gamma}_1^{*\top})^\top / \tau_1^2 \in \mathbb{R}^p$. Then, $\boldsymbol{\theta}_1$ satisfies

$$\boldsymbol{\theta}_1^\top \boldsymbol{\Sigma} = (1, -\boldsymbol{\gamma}_1^{*\top})\,\mathrm{E}\{\boldsymbol{U}\boldsymbol{U}^\top\}/\tau_1^2 = (1, \mathbf{0}_{p-1}^\top).$$

Hence, we have

$$(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^\top = \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Theta}.$$

We can directly bound

$$\|\boldsymbol{\gamma}_j^*\|_1 = \|\boldsymbol{\theta}_j / \Theta_{j,j}\|_1 - 1 \leq K_\gamma - 1 < K_\gamma.$$

By (D4), the minimal eigenvalue of $\boldsymbol{\Sigma}$ is at least ρ_*. We obtain through a spectral decomposition that the maximal eigenvalue of $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ is at most ρ_*^{-1}. Hence, we have

$$\tau_j^2 = \left(\mathbf{e}_j^\top \boldsymbol{\Theta} \mathbf{e}_j\right)^{-1} \geq \rho_*$$

and

$$\|\boldsymbol{\Theta}\|_1 \leq \max_{j=1,\ldots,p} \|\boldsymbol{\theta}_j / \Theta j, j\| \max_{j=1,\ldots,p} |\Theta_{j,j}| \leq K_\gamma / \rho_*.$$

□

Lemma B.8. *Denote*

$$\boldsymbol{\Delta}^{(l)}(t) = \mathbf{S}^{(l)}(t, \boldsymbol{\beta}^o) - \widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta}^o) \tag{B.6}$$

for $\mathbf{S}^{(l)}(t, \boldsymbol{\beta}^o)$ *in* (4) *and* $\widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta}^o)$ *in* (32). *Under (D1) - (D3) and (D5),*

(i) $\sup_{t\in[0,t^*]} \|\Delta^{(0)}(t)\|_{\max} = O_p\left(\sqrt{\log(n)/n}\right)$; $\sup_{l=1,2} \sup_{t\in[0,t^*]} \|\boldsymbol{\Delta}^{(l)}(t)\|_{\max}$, $\sup_{t\in[0,t^*]} \|\bar{\mathbf{Z}}(t, \boldsymbol{\beta}^o) - \widetilde{\mathbf{Z}}(t, \boldsymbol{\beta}^o)\|_\infty$, $\sup_{t\in[0,t^*]} \|\widetilde{\mathbf{Z}}(t, \boldsymbol{\beta}^o) - \boldsymbol{\mu}(t)\|_\infty$ *and* $\sup_{t\in[0,t^*]} \|\bar{\mathbf{Z}}(t, \boldsymbol{\beta}^o) - \boldsymbol{\mu}(t)\|_\infty$ *are all* $O_p\left(\sqrt{\log(p)/n}\right)$;

(ii) Define $\Delta_i(t) = \{\omega_i(t) - I(C_i > t)\}Y_i(t)$. *Let* $\boldsymbol{\phi}(\mathbf{Z})$ *be a differentiable operator* $\mathbb{R}^p \mapsto \mathbb{R}^q$ *uniformly bounded by* $K_\phi \asymp 1$ *with* $\|\nabla\boldsymbol{\phi}(\mathbf{Z})\|_1 < L_h \asymp 1$, *and* $\mathbf{g}(t)$ *be a* $\mathcal{F}_{t-}^*$ *adapted process in* $\mathbb{R}^{q'}$ *with bound* $\|\mathbf{g}(t)\|_{\max} = K_g \asymp 1$. *Whenever* $qq' = p$, *we have*

$$\left\| n^{-1/2} \sum_{i=1}^n \int_0^{t^*} n^{-1} \sum_{j=1}^n \Delta_j(t) \boldsymbol{\phi}(\mathbf{Z}_j(t)) \mathbf{g}(t)^\top I(C_i \geq t) dM_i^1(t) \right\|_{\max} = o_p(1); \tag{B.7}$$

(iii) for any $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p$, $\sup_{t\in[0,t^*]} \|\bar{\mathbf{Z}}(t, \boldsymbol{\beta}^o) - \bar{\mathbf{Z}}(t, \widetilde{\boldsymbol{\beta}})\|_\infty = O_p(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1)$; *if* $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = o_p(1)$,

$$\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} \left| \frac{e^{\boldsymbol{\beta}^{o\top} \mathbf{Z}_i(t)}}{S^{(0)}(t, \boldsymbol{\beta}^o)} - \frac{e^{\widetilde{\boldsymbol{\beta}}^\top \mathbf{Z}_i(t)}}{S^{(0)}(t, \widetilde{\boldsymbol{\beta}})} \right| = O_p(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1).$$

Proof of Lemma 6.

Since $\omega_i(t)Y_i(t) \neq I(C_i \geq t)Y_i(t)$ implies $\epsilon_i > 1$ thus $N_i^1(t^*) = 0$, we have the equivalence $dN_i^o(t) = \omega_i(t)dN_i^1(t) = I(C_i \geq t)dN_i^1(t)$. Recall for the following calculation that

$$\begin{aligned}
&\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o) = n^{-1}\sum_{i=1}^n \omega_i(t)Y_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l},\\
&\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o) = n^{-1}\sum_{i=1}^n I(C_i \geq t)Y_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l},\\
&\boldsymbol{\Delta}^{(l)}(t) = \mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o) - \widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o),\\
&\mathrm{E}\{\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)\} = \mathrm{E}\{\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o)\} = \mathbf{s}^{(l)}(t,\boldsymbol{\beta}^o)\\
&\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o) = \mathbf{S}^{(1)}(t,\boldsymbol{\beta}^o)/S^{(0)}(t,\boldsymbol{\beta}^o), \quad \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o) = \widetilde{\mathbf{S}}^{(1)}(t,\boldsymbol{\beta}^o)/\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o),\\
&\boldsymbol{\mu}(t) = \mathbf{s}^{(1)}(t,\boldsymbol{\beta}^o)/s^{(0)}(t,\boldsymbol{\beta}^o), \quad Y_i(t) = 1 - N_i^1(t-)\\
\text{and } &M_i^1(t) = N_i^1(t) - \int_0^t Y_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)}h_0^1(u)du.
\end{aligned}$$

We decompose

$$\begin{aligned}
\sqrt{n}\dot{\mathbf{m}}(\boldsymbol{\beta}^o) =& n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \left\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\right\} dN_i^o(t)\\
=& n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \left\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\right\} \omega_i(t)dM_i^1(t)\\
=& n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \left\{\boldsymbol{\mu}(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\right\} I(C_i \geq t)dM_i^1(t)\\
&+ n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \left\{\widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o) - \bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\right\} I(C_i \geq t)dM_i^1(t)\\
&+ n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \left\{\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o) - \boldsymbol{\mu}(t)\right\} \Delta^{(0)}(t)h_0^1(t)dt\\
&+ n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \left\{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\right\} \omega_i(t)dM_i^1(t)\\
\triangleq& I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

Notice that I_1 is a $\mathcal{F}_t^*$ martingale. We have $\|\boldsymbol{\mu}(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\|_\infty = O_p(\sqrt{\log(p)/n})$ from Lemma B.8(i). Hence, we can apply Lemma A.2(ii) to get $\|I_1\|_\infty = \sqrt{n}O_p(\sqrt{\log(p)/n}^2) = o_p(1)$.

We further decompose I_2 into 3 terms

$$\begin{aligned}
&- n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \frac{\boldsymbol{\Delta}^{(1)}(t)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)} I(C_i \geq t)dM_i^1(t) - n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \frac{\Delta^{(0)}(t)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\boldsymbol{\mu}(t) I(C_i \geq t)dM_i^1(t)\\
&+ n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \frac{\Delta^{(0)}(t)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\{\boldsymbol{\mu}(t) - \bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\} I(C_i \geq t)dM_i^1(t)\\
\triangleq& I_2' + I_2'' + I_2'''.
\end{aligned}$$

By (D1) and (D3), each $M_i^1(t)$ has one jump at observed event time and $e^{K_b}K_h-$Lipschitz elsewhere. Since the $\{C_i, T_i^1 : i = 1, \ldots, n\}$ is a set of independent continuous random variables, there is no tie among them with probability one. Hence, we may modify the integrand in I_2' and I_2'' at observed censoring times without changing the integral. Replacing $\boldsymbol{\Delta}^{(l)}(t)$ with $n^{-1}\sum_{j=1}^n \Delta_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}$, we can apply Lemma B.8(ii) to get that $\|I_2'\|_\infty$ and $\|I_2''\|_\infty$ are both $o_p(1)$.

The total variation of $M_i^1(t)$ is at most $\max\{1, e^{K_b}K_h t^*\} \asymp 1$. By Lemma B.8(i), $\|\boldsymbol{\Delta}^{(0)}(t)\{\boldsymbol{\mu}(t) - \bar{\mathbf{Z}}(t, \boldsymbol{\beta}^o)\}\|_\infty = O_p(\sqrt{\log(n)\log(p)}/n)$. Hence, we obtain $\|I_2'''\|_\infty = O_p(\sqrt{\log(n)\log(p)}/n) = o_p(1)$. Similarly, we obtain $\|I_3\|_\infty = O_p(\sqrt{\log(n)\log(p)}/n) = o_p(1)$.

Besides the one in Lemma B.6, $\omega_i(t) - \widetilde{\omega}_i(t)$ has another martingale representation. Denote the Nelson-Aalen estimator

$$\widehat{H}^c(t) = \sum_{i=1}^n \int_0^t \frac{I(X_i \geq u)}{\sum_{j=1}^n I(X_j \geq u)} dN_i^c(u).$$

We have a $\mathcal{F}_t$ martingale

$$\overline{M^c}(t) = \widehat{H}^c(t) - \int_0^t h^c(u)du = \sum_{i=1}^n \int_0^t \frac{I(X_i \geq u)}{\sum_{j=1}^n I(X_j \geq u)} dM_i^c(u).$$

By Lemma A.2(i), $\sup_{t\in[0,t^*]} |\overline{M^c}(t)| = O_p\left(n^{-1/2}\right)$ For $t > X_i$ and $\delta_i\epsilon_i > 1$,

$$\omega_i(t) - \widetilde{\omega}_i(t) = -\widetilde{\omega}_i(t)\int_0^t I(u > X_i)d\overline{M^c}(u) + R_i(t)$$

with an error

$$R_i(t) = \frac{\widehat{G}(t)}{\widehat{G}(X_i)} - \exp\left\{\widehat{H}^c(X_i) - \widehat{H}^c(t)\right\} + \frac{G(t)}{G(X_i)}\left[e^{-\int_0^t I(u>X_i)d\overline{M^c}(u)} + \int_0^t I(u > X_i)d\overline{M^c}(u)\right].$$

It is the discrepancy between the Kaplan-Meier and the Nelson-Aalen plus a second order Tailer expansion remainder. We shall show that it is $O_p(1/n)$. Since

$$\left|\int_0^t I(u > X_i)d\overline{M^c}(u)\right| \leq 2\sup_{t\in[0,t^*]} |\overline{M^c}(t)| = O_p\left(n^{-1/2}\right),$$

the second order remainder

$$\left|e^{-\int_0^t I(u>X_i)d\overline{M^c}(u)} + \int_0^t I(u > X_i)d\overline{M^c}(u)\right| = O_p(1/n).$$

Under (D2), $\{\sum_{i=1}^n I(X_i \geq t)\}^{-1} \leq \{\sum_{i=1}^n I(X_i \geq t^*)\}^{-1} = O_p(1/n)$. Let c_k be an observed censoring time. The increment in $-\log(\widehat{G}(t)) - \widehat{H}^c(t)$ at c_k is a second order remainder

$$\log\left(1 - \frac{1}{\sum_{i=1}^n I(X_i \geq c_k)}\right) - \frac{1}{\sum_{i=1}^n I(X_i \geq c_k)} = O_p\left(n^{-2}\right).$$

Hence, $\sup_{t\in[0,t^*]} |-\log(\widehat{G}(t)) - \widehat{H}^c(t)| = O_p(1/n)$. Applying the Mean Value Theorem, we obtain $\sup_{t\in[0,t^*]} |\widehat{G}(t) - \exp\{-\widehat{H}^c(t)\}| = O_p(1/n)$. Under (D2), $G(t) \geq G(t^*) > r_*$ and $-\log(G(t)) \leq -\log(G(t^*)) < -\log(r_*)$. We have shown that both $\widehat{G}(t)$ and $\widehat{H}^c(t)$ are uniformly $\sqrt{n}$ consistent. We obtain that $\widehat{G}(X_i)$ is bounded away from zero and $\widehat{H}^c(t)$ is bounded with probability tending to one. Putting these together, we obtain $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]} |R_i(t)| = O_p(1/n)$.

Define

$$\widetilde{\mathbf{q}}(t) = n^{-1}\sum_{i=1}^{n} I(t \geq X_i)\int_t^{t^*} \{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\}\,\widetilde{\omega}_i(u) dM_i^1(u),$$

$\widehat{\pi}(t) = n^{-1}\sum_{i=1}^n I(X_i \geq t)$ and $\mathbf{q}(t) = \mathrm{E}\{\widetilde{\mathbf{q}}(t)\}$, $\pi(t) = \mathrm{E}\{\widehat{\pi}(t)\}$. We write I_4 as i.i.d. sum plus error through integration by parts,

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\,\widetilde{\omega}_i(t) dM_i^1(t) + n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\,\{\omega_i(t) - \widetilde{\omega}_i(t)\} dM_i^1(t)\\
=&n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\,\widetilde{\omega}_i(t) dM_i^1(t) + n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\,R_i(t) dM_i^1(t)\\
&- n^{-1/2}\sum_{k=1}^{n}\int_0^{t^*} \frac{\mathbf{q}(t)}{\pi(t)} I(X_k \geq u) dM_k^c(t) + n^{-1/2}\sum_{k=1}^{n}\int_0^{t^*} \frac{\mathbf{q}(t)}{\widehat{\pi}(t)\pi(t)}\{\widehat{\pi}(t) - \pi(t)\} I(X_k \geq u) dM_k^c(t)\\
&+ n^{-1/2}\{\mathbf{q}(0) - \widetilde{\mathbf{q}}(0)\}\sum_{k=1}^{n}\int_0^{t^*} \frac{1}{\widehat{\pi}(t)} I(X_k \geq u) dM_k^c(t)\\
&- n^{-1/2}\sum_{k=1}^{n}\int_0^{t^*} \frac{\{\mathbf{q}(0) - \mathbf{q}(t) - \widetilde{\mathbf{q}}(0) + \widetilde{\mathbf{q}}(t)\}}{\widehat{\pi}(t)} I(X_k \geq u) dM_k^c(t)\\
\triangleq&I_4^{(1)} + I_4^{(2)} + I_4^{(3)} + I_4^{(4)} + I_4^{(5)} + I_4^{(6)}.
\end{aligned}$$

$I_4^{(1)} + I_4^{(3)}$ is already a sum of i.i.d.. We have shown that $\sup_{t\in[0,t^*]} |R_i(t)| = O_p(1/n)$. Hence, we have $\|I_4^{(2)}\|_\infty = O_p\left(n^{-1/2}\right) = o_p(1)$. $I(t \geq X_i)\int_t^{t^*} \{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\}\,\widetilde{\omega}_i(u) dM_i^1(u)$ is uniformly bounded by $K(K_h t^* + 1)$. It has at most one jump and is KK_h-Lipschitz elsewhere. Hence, we can apply Lemma A.1(ii) to get $\sup_{t\in[0,t^*]} \|\mathbf{q}(t) - \widetilde{\mathbf{q}}(t)\|_\infty = O_p(\sqrt{\log(p)/n})$ and $\sup_{t\in[0,t^*]} |\pi(t) - \widehat{\pi}(t)| = O_p(\sqrt{\log(n)/n})$. Notice that $I_4^{(4)}$, $I_4^{(6)}$ and $n^{-1}\sum_{k=1}^n \int_0^{t^*} \widehat{\pi}(t)^{-1} I(X_k \geq u) dM_k^c(t)$ in $I_4^{(5)}$ are all $\mathcal{F}_t$ martingales. We may apply Lemmas A.2(i) and A.2(ii) to obtain $I_4^{(4)} = O_p(\sqrt{\log(n)\log p/n}) = o_p(1)$, $I_4^{(5)} = O_p(\sqrt{\log p/n}) = o_p(1)$ and $I_4^{(6)} = O_p(\log p/\sqrt{n}) = o_p(1)$.

By Lemma 3, we can bound the l_1 norm of $\mathbf{c}^\top\boldsymbol{\Theta}$ by

$$\|\mathbf{c}^\top\boldsymbol{\Theta}\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{p} |c_i||\Theta_{ij}| \leq \sum_{i=1}^{p} |c_i| K_\gamma/\rho_* = K_\gamma/\rho_*.$$

Finally, we write $\mathbf{c}^\top\boldsymbol{\Theta}\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ as i.i.d. sum

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\mathbf{c}^\top\boldsymbol{\Theta}\left[\int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\,\widetilde{\omega}_i(t) dM_i^1(t) - \int_0^{t^*} \frac{\mathbf{q}(t)}{\pi(t)} I(X_i \geq u) dM_i^c(t)\right] + o_p(1)\\
\triangleq&n^{-1/2}\sum_{i=1}^{n}\mathbf{c}^\top\boldsymbol{\Theta}\{\boldsymbol{\eta}_i - \boldsymbol{\psi}_i\} + o_p(1).
\end{aligned}$$

We have $\mathrm{E}\{\mathbf{c}^\top\boldsymbol{\Theta}\boldsymbol{\eta}_i\} = 0$ because of its martingale structure. We show $\mathrm{E}\{\mathbf{c}^\top\boldsymbol{\Theta}\boldsymbol{\psi}_i\} = 0$ again by

introducing its martingale proxy

$$\begin{aligned}\mathrm{E}\{\mathbf{c}^\top\boldsymbol{\Theta}\boldsymbol{\psi}_i\} =& \mathrm{E}\left[\int_0^{t^*}\mathbf{c}^\top\boldsymbol{\Theta}\left\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\right\}I(C_i\geq t)dM_i^1(t)\right]\\ &+\mathrm{E}\left[\int_0^{t^*}\mathbf{c}^\top\boldsymbol{\Theta}\left\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\right\}\mathrm{E}\{\widetilde{\omega}_i(t)-I(C_i\geq t)|T_i,\mathbf{Z}_i(\cdot)\}dM_i^1(t)\right].\end{aligned}$$

The first term above is zero because of the martingale structure. The second term is zero because the IPW weights satisfy $\mathrm{E}\{\widetilde{\omega}_i(t)-I(C_i\geq t)|T_i,\mathbf{Z}_i(\cdot)\}=0$. Each $\mathbf{c}^\top\boldsymbol{\Theta}\{\boldsymbol{\psi}_i-\boldsymbol{\eta}_i\}$ is mean zero and bounded by $K_\gamma/\rho_*K(1+K_ht^*)+K_\gamma/\rho_*K(1+K_ht^*)(1+K_ct^*)2/r_*$ with probability equaling one. The variance $\mathbf{c}^\top\boldsymbol{\Theta}\boldsymbol{\mathcal{V}}\boldsymbol{\Theta}\mathbf{c}$ has a bounded and non-degenerating limit ν^2. Hence, $\{\mathbf{c}^\top\boldsymbol{\Theta}(\boldsymbol{\psi}_i-\boldsymbol{\eta}_i):i=1,\ldots,n\}$ satisfies the Lindeberg condition.

By Lindeberg-Feller CLT,

$$\sqrt{n}\frac{\mathbf{c}^\top\boldsymbol{\Theta}\dot{\mathbf{m}}(\boldsymbol{\beta}^o)}{\sqrt{\mathbf{c}^\top\boldsymbol{\Theta}\boldsymbol{\mathcal{V}}\boldsymbol{\Theta}\mathbf{c}}}=\frac{\mathbf{c}^\top\boldsymbol{\Theta}\sum_{i=1}^n\{\boldsymbol{\eta}_i-\boldsymbol{\psi}_i\}}{\sqrt{n\mathbf{c}^\top\boldsymbol{\Theta}\boldsymbol{\mathcal{V}}\boldsymbol{\Theta}\mathbf{c}}}+o_p(1)\xrightarrow{d}N(0,1).$$

□

Proof of Lemma 7.

We define

$$\widetilde{\boldsymbol{\eta}}_i=\int_0^{t^*}\{\mathbf{Z}_i(u)-\boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)d\widetilde{M}_i^1(u),$$

with

$$\widetilde{M}_i^1(t)=N_i^o(t)-n^{-1}\sum_{j=1}^n\int_0^t\frac{Y_i(u)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)}}{\widetilde{S}^{(0)}(u,\boldsymbol{\beta}^o)}dN_j^o(u).$$

Under (D1) and (D2), the total variation of $\widetilde{M}_i^1(t)$ is at most $1+2e^{2K_b}/r_*$ with probability tending to one by Lemma B.4*. The difference between $\widetilde{\boldsymbol{\eta}}_i$ and $\widehat{\boldsymbol{\eta}}_i$ is

$$\begin{aligned}\widehat{\boldsymbol{\eta}}_i-\widetilde{\boldsymbol{\eta}}_i=&n^{-1}\sum_{j=1}^n\int_0^{t^*}\{\mathbf{Z}_i(u)-\bar{\mathbf{Z}}(u,\widehat{\boldsymbol{\beta}})\}\omega_i(u)Y_i(u)\left\{\frac{e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)}}{\widetilde{S}^{(0)}(u,\boldsymbol{\beta}^o)}-\frac{e^{\widehat{\boldsymbol{\beta}}^\top\mathbf{Z}_i(u)}}{S^{(0)}(u,\widehat{\boldsymbol{\beta}})}\right\}dN_j^o(u)\\ &+\int_0^{t^*}\{\boldsymbol{\mu}(u)\widetilde{\omega}_i(u)-\bar{\mathbf{Z}}(u,\widehat{\boldsymbol{\beta}})\omega_i(u)\}d\widetilde{M}_i^1(u).\end{aligned}$$

By Lemmas B.6, B.8(i) and B.8(iii), $\sup_{i=1,\ldots,n}\|\widehat{\boldsymbol{\eta}}_i-\widetilde{\boldsymbol{\eta}}_i\|_\infty=O_p\left(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1+\sqrt{\log(p)/n}\right)$.

Then, we study

$$\boldsymbol{\eta}_i-\widetilde{\boldsymbol{\eta}}_i=n^{-1}\sum_{j=1}^n\int_0^{t^*}\{\mathbf{Z}_i(u)-\boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)I(C_j\geq u)dM_j^1(u).$$

We have the bound $\|\mathbf{Z}_i(u)-\boldsymbol{\mu}(u)\|_\infty\leq K$ from Lemma B.3. $\widetilde{\omega}_i(u)$ is not $\mathcal{F}_t^*$ measurable, but we can define a new filtration $\mathcal{F}_{i,t}^*=\sigma\{X_i,\delta_i,\epsilon_i,\mathbf{Z}_i(\cdot),I(C_j\geq u),N_j^1(u),\mathbf{Z}_j(\cdot):u\leq t,j\neq i\}$ for each i, such that

$$n^{-1}\sum_{j\neq i}\int_0^{t^*}\{\mathbf{Z}_i(u)-\boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)I(C_j\geq u)dM_j^1(u)=\boldsymbol{\eta}_i-\widetilde{\boldsymbol{\eta}}_i+O_p(1/n)$$

is a $\mathcal{F}^*_{i,t}$ martingale. Hence, we can apply Lemma A.2(i) to get

$$\Pr\left(\|\boldsymbol{\eta}_i - \widetilde{\boldsymbol{\eta}}_i\|_\infty \geq K(1+e^{K_b}K_h t^*)\sqrt{4\log(2np/\varepsilon)/n} + K(1+2e^{K_b}K_h t^*)/n\right) \leq \varepsilon/n.$$

Taking union bound, we get $\|\boldsymbol{\eta}_i - \widetilde{\boldsymbol{\eta}}_i\|_\infty = O_p(\sqrt{\log(p)/n})$. Hence, $\sup_{i=1,\ldots,n}\|\widehat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right)$.

Recall that $\widehat{\mathbf{q}}(t)$ and $\mathbf{q}(t)$ also take a similar form. We can likewise define

$$\widetilde{\mathbf{q}}(t) = n^{-1}\sum_{i=1}^n I(t > X_i)\int_t^{t^*}\{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)d\widetilde{M}^1_i(u)$$

and

$$\widetilde{\mathbf{q}}^*(t) = n^{-1}\sum_{i=1}^n I(t > X_i)\int_t^{t^*}\{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)dM^1_i(u).$$

By Lemmas B.6, B.8(i) and B.8(iii), we have

$$\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\|\widetilde{\mathbf{q}}(t) - \widehat{\mathbf{q}}(t)\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right).$$

By Lemma A.1(ii), $\sup_{t\in[0,t^*]}\|\widetilde{\mathbf{q}}^*(t) - \mathbf{q}(t)\| = O_p\left(\sqrt{\log(p)/n}\right)$. We only need to find the rate for

$$\widetilde{\mathbf{q}}^*(t) - \widetilde{\mathbf{q}}(t) = n^{-1}\sum_{i=1}^n I(t > X_i)n^{-1}\sum_{j=1}^n\int_t^{t^*} n^{-1}\sum_{i=1}^n\{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)I(C_j \geq u)dM^1_j(u).$$

We repeat the trick for $\boldsymbol{\eta}_i - \widetilde{\boldsymbol{\eta}}_i$. Applying Lemma A.2(ii) to the $\mathcal{F}^*_{i,t}$ martingale

$$\mathbf{M}^q_i(t) = n^{-1}\sum_{j\neq i}\int_0^t n^{-1}\sum_{i=1}^n\{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)I(C_j \geq u)dM^1_j(u)$$

and obtain $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\|\mathbf{M}^q_i(t)\|_\infty = O_p(\sqrt{\log(p)/n})$. Hence,

$$\sup_{t\in[0,t^*]}\|\widetilde{\mathbf{q}}^*(t) - \widetilde{\mathbf{q}}(t)\|_\infty \leq 2\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\|\mathbf{M}^q_i(t)\|_\infty + O_p(1/n) = O_p(\sqrt{\log(p)/n}).$$

Putting the rates together, we have $\sup_{t\in[0,t^*]}\|\widehat{\mathbf{q}}(t) - \mathbf{q}(t)\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right)$.

We can directly obtain $\sup_{t\in[0,t^*]}|\widehat{\pi}(t) - \pi(t)| = O_p\left(\sqrt{\log(n)/n}\right)$ from Lemma A.1(ii). Define

$$\widetilde{\boldsymbol{\psi}}_i = \int_0^{t^*}\frac{\mathbf{q}(t)}{\pi(t)}d\widehat{M}^c_i(t)$$

The total variation of $\widehat{M}^c_i(t)$ is at most $1 + 2/r_*$ with probability tending to one by Lemma B.4*. Using the results so far, we have $\sup_{i=1,\ldots,n}\|\widehat{\boldsymbol{\psi}}_i - \widetilde{\boldsymbol{\psi}}_i\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right)$. The remainder

$$\boldsymbol{\psi}_i - \widetilde{\boldsymbol{\psi}}_i = n^{-1}\sum_{j=1}^n\int_0^{t^*}\frac{\mathbf{q}(t)}{\pi(t)}I(C_i \geq t)I(X_j \geq t)dM^c_j(t)$$

is a $\mathcal{F}_t$ martingale. We can put the n martingales in $\mathbb{R}^p$ into a $\mathbb{R}^{np}$ vector and apply Lemma A.2(i),

$$\sup_{i=1,\ldots,n} \|\boldsymbol{\psi}_i - \widetilde{\boldsymbol{\psi}}_i\|_\infty = O_p\left(\sqrt{\log(np)/n}\right) = O_p\left(\sqrt{\log(p)/n}\right).$$

Therefore, we get $\sup_{i=1,\ldots,n} \|\boldsymbol{\psi}_i - \widehat{\boldsymbol{\psi}}_i\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right)$.

Finally, we decompose

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}}\|_{\max} \leq & n^{-1}\sum_{i=1}^n \|\widehat{\boldsymbol{\eta}}_i + \widehat{\boldsymbol{\psi}}_i\|_\infty \|\widehat{\boldsymbol{\eta}}_i + \widehat{\boldsymbol{\psi}}_i - \boldsymbol{\eta}_i - \boldsymbol{\psi}_i\|_\infty \\ & + n^{-1}\sum_{i=1}^n \|\widehat{\boldsymbol{\eta}}_i + \widehat{\boldsymbol{\psi}}_i - \boldsymbol{\eta}_i - \boldsymbol{\psi}_i\|_\infty \|\boldsymbol{\eta}_i + \boldsymbol{\psi}_i\|_\infty \\ & + \left\| n^{-1}\sum_{i=1}^n (\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)^\top - \boldsymbol{\mathcal{V}} \right\|_{\max}.\end{aligned}$$

We have shown that $\sup_{i=1,\ldots,n} \|\widehat{\boldsymbol{\eta}}_i + \widehat{\boldsymbol{\psi}}_i - \boldsymbol{\eta}_i - \boldsymbol{\psi}_i\|_\infty = o_p(1)$. Moreover, $\sup_{i=1,\ldots,n} \|\widehat{\boldsymbol{\eta}}_i + \widehat{\boldsymbol{\psi}}_i\|_\infty$ is $O_p(1)$ by Lemmas B.3 and B.4*. In addition, we observe that $n^{-1}\sum_{i=1}^n (\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)^\top$ is an average of i.i.d. terms whose expectation is defined as $\boldsymbol{\mathcal{V}}$. By Lemmas B.3 and B.4*, we have the uniform maximal bound

$$\sup_{i=1,\ldots,n} \|(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)^\top\|_{\max} = \sup_{i=1,\ldots,n} \|(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)\|_\infty^2$$

is also $O_p(1)$. We finish the proof by applying Hoeffding (1963) inequality to the last term in the decomposition above, $\left\|n^{-1}\sum_{i=1}^n (\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)^\top - \boldsymbol{\mathcal{V}}\right\|_{\max}$.

□

Lemma B.9. *On the event* $\Omega_\gamma := \left\{\left\|\nabla_\gamma \boldsymbol{\Gamma}_j\left(\boldsymbol{\gamma}_j^*, \widehat{\boldsymbol{\beta}}\right)\right\|_\infty \leq (\xi_j - 1)\lambda_j/(\xi_j + 1), \forall j = 1, \ldots, p\right\}$, *we have under (D4)*

(i) the estimation error $\widetilde{\boldsymbol{\gamma}}_j := \widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*$ *belongs to the cone*

$$\mathcal{C}_j(\xi_j, \mathcal{O}_j) := \{\mathbf{v} \in \mathbb{R}^{p-1} : \|\mathbf{v}_{\mathcal{O}_j^c}\|_1 \leq \xi_j \|\mathbf{v}_{\mathcal{O}_j}\|_1\} \tag{B.8}$$

(ii) and $\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1 \leq \{s_j\lambda_j(\xi_j + 1)\}/\{2\kappa_j(\xi_j, \mathcal{O}_j)^2\}$, *with compatibility factor*

$$\kappa_j(\xi_j, \mathcal{O}_j) = \sup_{0 \neq \mathbf{g} \in \mathcal{C}_j(\xi_j, \mathcal{O}_j)} \frac{\sqrt{s_j \mathbf{g}^\top \nabla_\gamma^2 \boldsymbol{\Gamma}(\boldsymbol{\gamma}^*, \widehat{\boldsymbol{\beta}})\mathbf{g}}}{\|\mathbf{g}_{\mathcal{O}_j}\|_1} \tag{B.9}$$

for all $j = 1, \ldots, p$.

Lemma B.10. *Under (D1)-(D5),* $\max_{j=1,\ldots,p} \left\|\nabla_\gamma \boldsymbol{\Gamma}_j\left(\boldsymbol{\gamma}_j^*, \widehat{\boldsymbol{\beta}}\right)\right\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right)$.

Lemma B.11. *Under (D1)-(D5),*

(i) $\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\max} = O_p\left(s_o\sqrt{\log(p)/n}\right)$;

(ii) for any $\widetilde{\boldsymbol{\beta}}$ such that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = o_p(1)$,

$$\left\| -\ddot{\mathbf{m}}(\widetilde{\boldsymbol{\beta}}) - \boldsymbol{\Sigma} \right\|_{\max} = O_p\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n} \right).$$

Lemma B.12. *Under (D1)-(D5), setting $\xi_{\max} = \max_{j=1,\ldots,p} \xi_j \asymp 1$, we have*

$$\Pr\left(\inf_j \kappa_j(\xi_j, \mathcal{O}_j)^2 \geq \rho_*/2 \right) \to 1.$$

Proof of Lemma 4.

By Lemma B.10, we may choose $\xi_1 = \cdots = \xi_p = 2$ and $\lambda_1 = \cdots = \lambda_p = \lambda_\varepsilon \asymp O_p(s_o\sqrt{\log(p)/n})$ such that Ω_γ defined in Lemma B.9 occurs with probability $1 - \varepsilon$. Then, we establish the oracle inequality by Lemma B.9,

$$\Pr\left(\max_{j=1,\ldots,p} \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1/s_j \leq \frac{2\lambda_\varepsilon}{\rho_*} \right) \geq \Pr\left(\min_{j=1,\ldots,p} \kappa_j(\xi_j, \mathcal{O}_j)^2 \geq \rho_*/2 \right) - \varepsilon.$$

We have shown that $\Pr\left(\min_{j=1,\ldots,p} \kappa_j(\xi_j, \mathcal{O}_j)^2 \geq \rho_*/2\right)$ tends to one in Lemma B.12. Hence, $\max_{j=1,\ldots,p} \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1 = O_p\left(s_o s_{\max}\sqrt{\log(p)/n}\right)$.

Define according to (49) $\boldsymbol{U}_i = \int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\} dN_i^o(t)$. By Lemma B.3, $\sup_{i=1,\ldots,n} \|\boldsymbol{U}_i\|_\infty \leq K$. We introduce

$$\widetilde{\Gamma}_j(\boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^n \{U_j - \boldsymbol{U}_{i,-j}^\top \boldsymbol{\gamma}_j^*\} = n^{-1} \sum_{i=1}^n \int_0^{t^*} \{Z_{ij}(t) - \mu_j(t) - \boldsymbol{\gamma}^\top \mathbf{Z}_{i,-j}(t) + \boldsymbol{\gamma}^\top \boldsymbol{\mu}_{-j}(t)\}^2 dN_i^o(t)$$

and decompose

$$\widehat{\tau}_j^2 - \tau_j^2 = \Gamma_j(\widehat{\boldsymbol{\gamma}}_j, \widehat{\boldsymbol{\beta}}) - \widetilde{\Gamma}_j(\boldsymbol{\gamma}_j^*) + \widetilde{\Gamma}_j(\boldsymbol{\gamma}_j^*) - \bar{\Gamma}_j(\boldsymbol{\gamma}_j^*).$$

$\Gamma_j(\widehat{\boldsymbol{\gamma}}_j, \widehat{\boldsymbol{\beta}}) - \widetilde{\Gamma}_j(\boldsymbol{\gamma}_j^*) = O_p\left(s_o s_j \sqrt{\log(p)/n}\right)$ by the results from Theorem 1*, Lemma B.8 and first part of this Lemma. Apparently, $\widetilde{\Gamma}_j(\boldsymbol{\gamma}_j^*)$ is the average of i.i.d. terms. The expectation of the summands in $\widetilde{\Gamma}_j(\boldsymbol{\gamma}_j^*)$ is defined as $\bar{\Gamma}_j(\boldsymbol{\gamma}_j^*)$ in (49). Hence, we finish the proof by applying Hoeffding (1963).

Along with Lemma 3, we can prove with the previous results in this Lemma, $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1 = O_p\left(s_o s_{\max}\sqrt{\log(p)/n}\right)$. □

Proof of Lemma 5.

We decompose

$$\sqrt{n}\mathbf{c}^\top \left\{ \boldsymbol{\Theta}\dot{\mathbf{m}}(\boldsymbol{\beta}^o) + \boldsymbol{\beta}^o - \widehat{\mathbf{b}} \right\} = \sqrt{n}\mathbf{c}^\top \{\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\}\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}}) + \sqrt{n}\mathbf{c}^\top \boldsymbol{\Theta}\{\dot{\mathbf{m}}(\boldsymbol{\beta}^o) - \dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}})\} + \sqrt{n}\mathbf{c}^\top(\boldsymbol{\beta}^o - \widehat{\boldsymbol{\beta}}). \tag{B.10}$$

By Lemma 4, $\|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_1 = O_p(s_o s_{\max}\sqrt{\log(p)/n})$. Each summand in $\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}})$ is the integral of $\mathbf{Z}_i(t)$ minus a weighted average $\bar{\mathbf{Z}}(t, \widehat{\boldsymbol{\beta}})$ over a counting measure $dN_i^o(t)$. By the KKT condition and Theorem 1*, $\|\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}})\|_\infty \asymp \lambda \asymp O(\sqrt{\log(p)/n})$. Putting these together, we obtain

$$\sqrt{n}|\mathbf{c}^\top \{\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\}\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}})| \leq \sqrt{n}\|\mathbf{c}\|_1 \|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_1 \|\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}})\|_\infty = O_p\left(s_o s_{\max} \log(p)/\sqrt{n}\right) = o_p(1).$$

By the KKT condition and Theorem 1, $\|\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}})\| \leq \lambda \asymp n^{-(1/2-d)}$. Hence, the first term in (B.10) is $o_p(1)$. Like in the proof of Lemma 6, we have $\|\mathbf{c}^\top \boldsymbol{\Theta}\|_1 \leq \|\mathbf{c}\|_1 \|\boldsymbol{\Theta}\|_1 \leq K_\gamma/r_*$ from Lemma 3.

Define $\boldsymbol{\beta}_r = \boldsymbol{\beta}^o + r(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)$. Applying mean value theorem to $h(r) = \mathbf{c}^\top \boldsymbol{\Theta} \dot{\mathbf{m}}(\beta_r)$, we get

$$\mathbf{c}^\top \boldsymbol{\Theta} \dot{\mathbf{m}}(\boldsymbol{\beta}^o) - \mathbf{c}^\top \boldsymbol{\Theta} \dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}}) = -h'(\widetilde{r}) = -\mathbf{c}^\top \boldsymbol{\Theta} \ddot{\mathbf{m}}(\boldsymbol{\beta}_{\widetilde{r}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)$$

for some $\widetilde{r} \in [0,1]$. By Theorem 1*, we have $\|\boldsymbol{\beta}_{\widetilde{r}} - \boldsymbol{\beta}^o\|_1 = \widetilde{r}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = O_p\left(s_o\sqrt{\log(p)/n}\right)$. By Lemma B.11(ii), $\|-\ddot{\mathbf{m}}(\boldsymbol{\beta}_{\widetilde{r}}) - \boldsymbol{\Sigma}\|_{\max} = O_p\left(s_o\sqrt{\log(p)/n}\right)$. Along with Theorem 1* and Lemma 3, we have

$$\begin{aligned}\sqrt{n}|\mathbf{c}^\top \boldsymbol{\Theta}\{\dot{\mathbf{m}}(\boldsymbol{\beta}^o) - \dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}})\} + \mathbf{c}^\top(\boldsymbol{\beta}^o - \widehat{\boldsymbol{\beta}})| =& \sqrt{n}|\mathbf{c}^\top \boldsymbol{\Theta}\{\boldsymbol{\Sigma} + \ddot{\mathbf{m}}(\boldsymbol{\beta}_{\widetilde{r}})\}(\boldsymbol{\beta}^o - \widehat{\boldsymbol{\beta}})| \\ \le& \sqrt{n}\|\mathbf{c}\|_1\|\boldsymbol{\Theta}\|_1\|-\ddot{\mathbf{m}}(\boldsymbol{\beta}_{\widetilde{r}}) - \boldsymbol{\Sigma}\|_{\max}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \\ =& O_p\left(s_o^2\log(p)/\sqrt{n}\right).\end{aligned}$$

□

Proof of Theorem 2.

Be Lemmas 6 and 5, we have

$$\sqrt{n}\frac{\mathbf{c}^\top(\widehat{\mathbf{b}} - \boldsymbol{\beta}^o)}{\mathbf{c}^\top \boldsymbol{\Theta}\boldsymbol{\mathcal{V}}\boldsymbol{\Theta}^\top \mathbf{c}} = \sqrt{n}\frac{\boldsymbol{\Theta}\dot{\mathbf{m}}(\boldsymbol{\beta}^o)}{\mathbf{c}^\top \boldsymbol{\Theta}\boldsymbol{\mathcal{V}}\boldsymbol{\Theta}^\top \mathbf{c}} + o_p(1) \xrightarrow{d} N(0,1).$$

In Lemma 7, we have shown that $\|\boldsymbol{\mathcal{V}}\|_{\max}$ is bounded by $K^2(1 + K_h e^{K_b t^*})^2\{1 + 2(1+K_c)/r_*\}^2$ with probability tending to one. In Lemma 3, we have shown that $\|\boldsymbol{\Theta}\|_1$ is bounded by K_γ/ρ_*. Then, we can apply Lemmas 7 and 4 to get

$$\begin{aligned}|\mathbf{c}^\top \boldsymbol{\Theta}\boldsymbol{\mathcal{V}}\boldsymbol{\Theta}^\top \mathbf{c} - \mathbf{c}^\top \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\mathcal{V}}}\widehat{\boldsymbol{\Theta}}^\top \mathbf{c}| \le& \|\mathbf{c}\|_1\|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_1\|\boldsymbol{\mathcal{V}}\|_{\max}\|\boldsymbol{\Theta}\|_1\|\mathbf{c}\|_1 \\ &+ \|\mathbf{c}\|_1\{\|\boldsymbol{\Theta}\|_1 + \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1\}\|\boldsymbol{\mathcal{V}} - \widehat{\boldsymbol{\mathcal{V}}}\|_{\max}\|\boldsymbol{\Theta}\|_1\|\mathbf{c}\|_1 \\ &+ \|\mathbf{c}\|_1\{\|\boldsymbol{\Theta}\|_1 + \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1\}\{\|\widehat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}}\|_{\max} + \|\boldsymbol{\mathcal{V}}\|_{\max}\}\|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_1\|\mathbf{c}\|_1 \\ =& 2O_p(\|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_1) + O_p(\|\boldsymbol{\mathcal{V}} - \widehat{\boldsymbol{\mathcal{V}}}\|_{\max}) = o_p(1).\end{aligned}$$

Note that we use the following fact

$$\|\mathbf{c}^\top \boldsymbol{\Theta}\|_1 = \sum_{j=1}^p |\sum_{i=1}^p c_i \Theta_{i,j}| \le \sum_{i=1}^p |c_i| \sum_{j=1}^p |\Theta_{i,j}| \le \|\mathbf{c}\|_1\|\boldsymbol{\Theta}\|_1.$$

□

C Proofs of Preliminary Results

Proof of Lemma B.3.

Notice all $a_i(t)$'s are nonnegative. Hence, $\sum_{i=1}^n |a_i(t)| = \sum_{i=1}^n a_i(t)$. We apply Hölder's inequality for each coordinate

$$\left|\left\{\frac{\sum_{i=1}^n a_i(t)\mathbf{Z}_i(t)^{\otimes l}}{\sum_{i=1}^n a_i(t)}\right\}_j\right| = \left|\sum_{i=1}^n \frac{a_i(t)}{\sum_{i=1}^n a_i(t)}\left\{\mathbf{Z}_i(t)^{\otimes l}\right\}_j\right| \le \frac{\sum_{i=1}^n |a_i(t)|}{|\sum_{i=1}^n a_i(t)|}\sup_{i=1,\ldots,n}\left|\left\{\mathbf{Z}_i(t)^{\otimes l}\right\}_j\right|.$$

Hence, the maximal norm of $\sum_{i=1}^n a_i(t)\mathbf{Z}_i(t)^{\otimes l}$ is bounded by $(K/2)^l$ under (39). Similar result can be achieved with the sum replaced by the expectation.

To apply the result above to $\mathbf{S}^{(l)}(t,\boldsymbol{\beta})/S^{(0)}(t,\boldsymbol{\beta})$, $\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta})/\widetilde{S}^{(0)}(t,\boldsymbol{\beta})$ and $\mathbf{s}^{(l)}(t,\boldsymbol{\beta})/s^{(0)}(t,\boldsymbol{\beta})$, we set $a_i(t)$ as $\omega_i(t)Y_i(t)e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)}$ and $I(C_i \ge t)Y_i(t)e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)}$. □

Proof of Lemma B.4.

Since $\{I(X_i \geq t^*), i = 1, \ldots, n\}$ are i.i.d. Bernoulli random variable, we may apply Hoeffding (1963) inequality for lower tail,

$$\Pr\left(n^{-1}\sum_{i=1}^{n} I(X_i \geq t^*) < \Pr(X_i \geq t^*) - x\right) \leq \exp(-2nx^2).$$

By (40), we can find lower bounds for the probability

$$\Pr(X_i \geq t^*) \geq \Pr(C_i \geq t^*, \infty > T_i^1 \geq t^*) = G(t^*)\,\mathrm{E}\{F_1(\infty; \mathbf{Z}_i) - F_1(t^*; \mathbf{Z}_i)\} \geq r_*/M.$$

We may relax the inequality at $x = r_*/(2M)$ to

$$\Pr\left(n^{-1}\sum_{i=1}^{n} I(X_i \geq t^*) < r_*/(2M)\right) \leq e^{-nr_*^2/(2M^2)}.$$

Because $I(C_i \geq t) \geq I(C_i \geq t^*)$ and $Y_i(t) \geq Y_i(t^*)$, $\widetilde{S}^{(0)}(t; M)$ is a lower bound for $\widetilde{S}^{(0)}(t)$. The summands in $\widetilde{S}^{(0)}(t; M)$ are i.i.d. uniformly bounded by M. Thus, we may apply Lemma A.1(i) with one-sided version,

$$\Pr\left(\sup_{k\in 1\ldots K_N} \widetilde{S}^{(0)}(t; M) < \mathrm{E}\{\widetilde{S}^{(0)}(t; M)\} - Mx - M/n\right) < ne^{-nx^2/2}.$$

By (C2), the expectation has a lower bound

$$\mathrm{E}\{\widetilde{S}^{(0)}(t; M)\} = G(t^*)\,\mathrm{E}\left[\{1 - F_1(t; \mathbf{Z}_i)\}\min\{M, e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\}\right] > r_*.$$

We relax the inequality at $x = (r_*/2 - 1/n)/M$,

$$\Pr\left(\sup_{k\in 1\ldots K_N} \widetilde{S}^{(0)}(t; M) < r_*\right) < ne^{-n(r_*-2/n)^2/(8M^2)}.$$

□

Proof of Lemma B.5.

Since $\epsilon_i > 1$ implies $T_i^1 = \infty$, the probability of observing a type-2 event conditioning on $\mathbf{Z}_i(\cdot)$ has an upper bound

$$\begin{aligned}\Pr\left(\epsilon_i > 1 | \mathbf{Z}_i(\cdot)\right) &= \exp\left\{-\int_0^\infty e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)} h_0^1(u)du\right\} \\ &\leq \exp\left\{-K_e x \int_0^\infty I\left(e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)} \geq K_e x\right) h_0^1(u)du\right\}.\end{aligned}$$

Hence, we may derive a bound for

$$\Pr\left(\delta_i\epsilon_i > 1, \sup_{t\in[0,t^*]} e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)} > K_e\right) \leq \Pr\left(\epsilon_i > 1 \Big| \sup_{t\in[0,t^*]} e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)} > K_e\right)$$

if we can bound $\int_0^\infty I\left(e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)} \geq K_e x\right) h_0^1(u)du$ away from zero with a certain x whenever $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t')} > K_e$ for some $t' \in [0, t^*]$.

Under (C2), there is an interval I' containing t' of length D in which $\mathbf{Z}_i(\cdot)$ has no jumps. The variation of linear predictor is bounded

$$\sup_{t\in I'} \left|\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t) - \boldsymbol{\beta}^o\mathbf{Z}_i(t')\right| \leq c_z L_z \|\boldsymbol{\beta}^o\|_\infty D.$$

So, the relative risk $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}$ is greater than $K_e \exp\{-c_z L_z \|\boldsymbol{\beta}^o\|_\infty D\}$ over I'. Hence, we get a lower bound for

$$\int_0^\infty I\left(e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)} \geq K_e \exp\{-c_z L_z \|\boldsymbol{\beta}^o\|_\infty D\}\right) h_0^1(u)du \geq D h_*.$$

We finish the proof by taking a union bound over $i = 1, \ldots, n$. □

Proof of Lemma B.6.

Recall that $M_i^c(t) = I(C_i \leq t) - \int_0^t I(C_i \geq u)h^c(u)du$ is a counting process martingale adapted to complete data filtration $\mathcal{F}_t$. By Kalbfleisch and Prentice (2002), the Kaplan-Meier estimator $\widehat{G}(t)$ has the martingale representation

$$M^G(t) = \frac{\widehat{G}(t)}{G(t)} - 1 = n^{-1}\sum_{i=1}^n \int_0^t \frac{\widehat{G}(u-)I(X_i \geq u)}{G(u)n^{-1}\sum_{j=1}^n I(X_j \geq u)} dM_i^c(u).$$

For $\delta_i\epsilon_i > 1$ and $t > X_i$,

$$\omega_i(t) - \widetilde{\omega}_i(t) = -\frac{\widehat{G}(t)}{\widehat{G}(X_i)}M^G(X_i) + \frac{G(t)}{G(X_i)}M^G(t),$$

so we will be able to establish a concentration result for the error from Kaplan-Meier

$$\left\| n^{-1}\sum_{i=1}^n \{\omega_i(t) - \widetilde{\omega}_i(t)\} Y_i(t) e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes l} \right\|_{\max} \leq 2K_{e,\varepsilon}(K/2)^l \sup_{t\in[0,t^*]} |M^G(t)|$$

if we first obtain a concentration result for $\sup_{t\in[0,t^*]} |M^G(t)|$. On event $n^{-1}\sum_{j=1}^n I(X_j \geq u) \geq r_*/(2M)$, the integrated functions are $\mathcal{F}_{t-}$-adapted with uniform bound $2(M/r_*)^2$. The hazard $h^c(t) \leq K_c$ by (C1). Hence, we may apply Lemma A.2(i) with $x = \sqrt{4\log(2/\varepsilon)/n}$ to obtain

$$\Pr\left(\sup_{t\in[0,t^*]} |M^G(t)| < 2(M/r_*)^2 \left\{(1 + K_c t^*)\sqrt{4\log(2/\varepsilon)/n} + K_c t^*/n\right\}\right) \leq \Pr(\Omega_{r_*} \cap \Omega_{e,\varepsilon}) - \varepsilon.$$

□

Proof of Lemma B.7.

A sharper inequality is available if $\mathbf{Z}_i$'s are not time-dependent. We may exploit the martingale structure of $\boldsymbol{\Delta}^{(l)}(t)/G(t)$. With general time-dependent covariates, we would decompose the approximation error $\boldsymbol{\Delta}^{(l)}(t)$ into two parts, the error from Kaplan-Meier estimate $\widehat{G}(t)$ and the error from missingness in C_i's among the type-2 events.

Define the indicator $\upsilon_i(t) = I(t > X_i)I(\delta_i\epsilon_i > 1)$. Since $\{\omega_i(t) - I(C_i \geq t)\}Y_i(t)$ is non-zero only when $\upsilon(t) = 1$, we may alternatively write

$$\boldsymbol{\Delta}^{(l)}(t) = n^{-1}\sum_{i=1}^{n}\{\omega_i(t) - I(C_i \geq t)\}\upsilon_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}.$$

We may use the upper bound $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\left|\upsilon_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\right| \leq K_{e,\varepsilon}$ on $\Omega_{e,\varepsilon}$. By Lemma B.6,

$$\left\|n^{-1}\sum_{i=1}^{n}\{\omega_i(t) - \widetilde{\omega}_i(t)\}\upsilon_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}\right\|_{\max} \leq K_{e,\varepsilon}(K/2)^l C^{\omega}_{n,p,\varepsilon}$$

on $\Omega_{e,\varepsilon}\cap\Omega_{KM,\varepsilon}$.

Define the error from missingness in C_i's among the type-2 events as

$$\widetilde{\boldsymbol{\Delta}}^{(l)}(t) = n^{-1}\sum_{i=1}^{n}\{\widetilde{\omega}_i(t) - I(C_i \geq t)\}\upsilon_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}.$$

Since $\mathrm{E}\{r_i(t)|T_i\} = G(t\wedge T_i)$, Fine and Gray (1999) has shown that

$$\mathrm{E}\{\widetilde{\omega}_i(t)|T_i\} = \mathrm{E}\{I(C_i \geq t)|T_i\} = G(t).$$

Applying tower property, we have $\mathrm{E}\left\{\widetilde{\boldsymbol{\Delta}}^{(l)}(t)\right\} = \mathbf{0}$. Hence, we can apply Lemma A.1(i) with $x = \sqrt{2\log(2np^l/\varepsilon)/n}$

$$\Pr\left(\sup_{k\in 1\ldots K_N}\left\|\widetilde{\boldsymbol{\Delta}}^{(l)}\left(T^1_{(k)}\right)\right\|_{\max} \leq K_{e,\varepsilon}(K/2)^l\left\{\sqrt{2\log(2np^l/\varepsilon)/n} + 1/n\right\}\right) \geq \Pr(\Omega_{r_*}\cap\Omega_{e,\varepsilon}) - \varepsilon.$$

This finishes the proof of the first result.

We prove the other result by decomposing the differences into terms with $\boldsymbol{\Delta}^{(l)}(t)$,

$$\frac{\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)} = \frac{1}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\boldsymbol{\Delta}^{(l)}(t) - \frac{\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\Delta^{(0)}(t).$$

$\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)/S^{(0)}(t,\boldsymbol{\beta}^o)$ is the weighted average of $\mathbf{Z}_i(t)^{\otimes l}$, so its maximal norm is bounded by $(K/2)^l$. On the event Ω_{r_*},

$$\left\|\frac{\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\right\|_{\infty} \leq \frac{2}{r_*}\|\boldsymbol{\Delta}^{(l)}(t)\|_{\infty} + \frac{K^l}{2^{l-1}r_*}|\Delta^{(0)}(t)|.$$

We can simply plug in the bounds and tail probabilities for $\Delta^{(0)}\left(T^1_{(k)}\right)$ and $\boldsymbol{\Delta}^{(1)}\left(T^1_{(k)}\right)$ in (B.4). □

Proof of Lemma B.4*.

Consider the event

$$\Omega_{r_*} = \left\{n^{-1}\sum_{i=1}^{n} I(X_i \geq t^*)I(\epsilon_i = 1) \geq r_*/2\right\}.$$

Each $I(X_i \geq t^*)I(\epsilon_i = 1)$ is i.i.d. with expectation $G(t^*)\,\mathrm{E}[\{F_1(\infty;\mathbf{Z}) - F_1(t^*;\mathbf{Z})\}]$. Applying Hoeffding (1963) under (47), we get that Ω_{r_*} occurs with probability $1 - e^{-nr_*^2}$.

Apparently, we have $I(X_i \geq t^*) \geq I(X_i \geq t^*)I(\epsilon_i = 1)$. Morevoer, $S^{(0)}(t,\boldsymbol{\beta}^o)$ and $\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)$ are both lower bounded by $n^{-1}\sum_{i=1}^n I(X_i \geq t^*)e^{-K_b}$. On Ω_{r_*}, $\sup_{t\in[0,t^*]}|n/\{\sum_{i=1}^n I(X_i \geq t^*)\}| \leq 2/r_*$ and

$$\max\left\{\sup_{t\in[0,t^*]}|S^{(0)}(t,\boldsymbol{\beta}^o)^{-1}|, \sup_{t\in[0,t^*]}|\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)^{-1}|\right\} \leq 2e^{K_b}/r_*.$$

□

Proof of Lemma B.8.

(i) By (D1) and (D2), we have $\left\|e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}\right\|_{\max} \leq (K/2)^l e^{K_b} \asymp 1$. Thus, all terms involved are bounded. Moreover, $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}$ jumps only at the jumps of $N_i^z(t)$ by (D3). Define the outer product of arrays $\mathbf{u} \in \mathbb{R}^{p_1\times\cdots\times p_d}$ and $\mathbf{v} \in \mathbb{R}^{q_1\times\cdots\times q_{d'}}$ as

$$\mathbf{u}\otimes\mathbf{v} \in \mathbb{R}^{p_1\times\cdots\times p_d\times q_1\times\cdots\times q_{d'}}, \quad (\mathbf{u}\otimes\mathbf{v})_{i_1,\ldots,i_{d+d'}} = \mathbf{u}_{i_1,\ldots,i_d} \times \mathbf{v}_{i_{d+1},\ldots,i_{d+d'}}.$$

Between two consecutive jumps of $N_i^z(t)$,

$$\begin{aligned}\left\|\frac{d}{dt}e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}\right\|_{\max} &= \left\|e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}\boldsymbol{\beta}^{o\top}\mathbf{d}_i^z(t) + I(l>0)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}l\mathbf{Z}_i(t)^{\otimes l-1}\otimes\mathbf{d}_i^z(t)\right\|_{\max}\\ &\leq e^{K_b}\{(K/2)^l L_z + I(l>1)(K/2)^{l-1}L_z\} \asymp 1.\end{aligned}$$

Hence, $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}$ satisfies the continuity condition for Lemma A.1(ii).

Like in Lemma B.7, we first replace $\omega_i(t)$ by $\widetilde{\omega}_i(t) = r_i(t)G(t)/G(X_i\wedge t)$. Denote $\widetilde{\boldsymbol{\Delta}}^{(l)}(t) = n^{-1}\sum_{i=1}^n\{\widetilde{\omega}_i(t) - I(C_i \geq t)\}Y_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}$. By Lemma B.6, $\sup_{t\in[0,t^*]}\|\boldsymbol{\Delta}^{(l)}(t) - \widetilde{\boldsymbol{\Delta}}^{(l)}(t)\|_{\max} = O_p\left(n^{-1/2}\right)$. Then, we apply Lemma A.1(ii) to the i.i.d. mean zero process $\widetilde{\boldsymbol{\Delta}}^{(l)}(t)$,

$$\sup_{t\in[0,t^*]}\|\widetilde{\boldsymbol{\Delta}}^{(l)}(t)\|_{\max} = O_p\left(\sqrt{\log(np^lK_z)/n}\right).$$

Similarly,

$$\sup_{t\in[0,t^*]}\|\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o) - \mathbf{s}^{(l)}(t,\boldsymbol{\beta}^o)\|_{\max} = O_p\left(\sqrt{\log(np^lK_z)/n}\right).$$

Finally, we extend to results to the quotients by decomposition

$$\frac{\mathbf{S}^{(1)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{\widetilde{\mathbf{S}}^{(1)}(t,\boldsymbol{\beta}^o)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)} = \frac{1}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\boldsymbol{\Delta}^{(l)}(t) - \frac{\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\Delta^{(0)}(t).$$

The denominators are bounded away from zero by Lemma B.4* by choosing $M = e^{K_b}$.

(ii) First, we show that $\Delta_i(t)$ is related to the martingales $M_i^c(t) = N_i^c(t) - \int_0^t\{1-N_i^c(u-)\}h^c(u)du$. $\Delta_i(t)$ is non-zero only after an observed type-2 event. To simplify notation, we define the indicator for non-zero $\Delta_i(t)$, $\upsilon_i(t) = r_i(t)Y_i(t)I(t > X_i) = I(\delta_i\epsilon_i > 1)I(t > X_i)$.

Denote the Nelson-Aalen type estimator for censoring cumulative hazard as

$$\widehat{H}^c(t) = \sum_{i=1}^n \int_0^t \Big\{ \sum_{j=1}^n I(X_j \geq u) \Big\}^{-1} I(X_i \geq u) dN_i^c(u).$$

Define $R_i(t) = \widehat{G}(t)/\widehat{G}(X_i) - 1 + \int_{X_i}^t \widehat{G}(u-) d\widehat{H}^c(u)/\widehat{G}(X_i)$. Let c_k and c_{k+1} be two consecutive observed censoring times greater than X_i. The increment $R_i(c_{k+1}) - R_i(c_k)$ is in fact

$$\frac{\widehat{G}(c_k)}{\widehat{G}(X_i)} \left\{ \frac{\sum_{j=1}^n I(X_j \geq c_{k+1}) - 1}{\sum_{j=1}^n I(X_j \geq c_{k+1})} - 1 + \frac{1}{\sum_{j=1}^n I(X_j \geq c_{k+1})} \right\} = 0.$$

For $t > X_i$, we have $R_i(t) = 0$. Thus,

$$\begin{aligned}
\Delta_i(t) =& \{\widehat{G}(t)/\widehat{G}(X_i) - 1 + N_i^c(t) - N_i^c(X_i) - R_i(t)\} \upsilon_i(t) \\
=& \int_{X_i}^t \upsilon_i(u) dM_i^c(u) - \int_{X_i}^t \omega_i(u-) \upsilon_i(u) \frac{\sum_{j=1}^n I(X_j \geq u) dM_j^c(u)}{\sum_{j=1}^n I(X_j \geq u)} + \\
&+ \int_{X_i}^t \{I(C_i \geq u) - \omega_i(u-)\} \upsilon_i(u) h^c(u) du. \qquad \text{(C.11)}
\end{aligned}$$

Notice $\upsilon_i(t)$ does not change beyond X_i if $C_i > X_i$, i.e. an event is observed. Since $h^c(u) \leq K_c < \infty$, we may modify the integrand at countable many points without changing the integral

$$\int_{X_i}^t \{I(C_i \geq u) - \omega_i(u-)\} \upsilon_i(u) h^c(u) du = -\int_{X_i}^t \Delta_i(u) h^c(u) du.$$

Hence, (C.11) gives an first order linear integral equation for $\Delta_i(u)$. The general solution to the related homogeneous problem

$$\Delta_i(t) = -\int_{X_i}^t \Delta_i(u) h^c(u) du, \quad \Delta_i(X_i) = 0$$

has only one unique solution $\Delta_i(t) = 0$. Thus, we only need to find one specific solution to (C.11). Define an integral operator $I \circ f = \int_{X_i}^t f(u) h^c(u) du$. Then, the solution to $f(t) = g(t) - I \circ f(t)$ can be written as

$$f(t) = (1 - I + I^2 - I^3 + \dots) \circ g(t) \triangleq e^{-I} \circ g(t).$$

By inductively using integration by parts, we are able to calculate

$$I^n \circ g(t) = \frac{1}{n!} \sum_{k=1}^n \binom{n}{k} (-1)^k H^c(t)^{n-k} \int_{X_i}^t H^c(u)^k dg(t).$$

Hence, the solution can be calculated as the series

$$\begin{aligned}
f(t) =& \sum_{n=1}^\infty (-1)^n I^n \circ g(t) = \sum_{n=1}^\infty \sum_{k=1}^n \frac{\{-H^c(t)\}^{n-k}}{(n-k)!} \int_{X_i}^t \frac{H^c(u)^k}{k!} dg(u) \\
=& \sum_{k=1}^\infty \int_{X_i}^t \frac{H^c(u)^k}{k!} dg(u) \sum_{n=k}^\infty \frac{\{-H^c(t)\}^{n-k}}{(n-k)!} = G(t) \int_{X_i}^t G(u)^{-1} dg(u).
\end{aligned}$$

Applying to (C.11), we get

$$\Delta_i(t) = G(t)\int_{X_i}^{t} G(u)^{-1} dM_i^{\Delta}(u),$$

with a $\mathcal{F}_t-$ martingale

$$M_i^{\Delta}(t) = \int_0^t I(C_i \geq u)\upsilon_i(u)dM_i^c(u) - \int_0^t \omega_i(u-)\upsilon_i(u)\frac{\sum_{j=1}^n I(X_j \geq u)dM_j^c(u)}{\sum_{j=1}^n I(X_j \geq u)}.$$

Now, we use the martingale structure to prove the Lemma. Denote the $\mathcal{F}_t^*$ martingale

$$\mathbf{M}^g(t) = n^{-1}\sum_{i=1}^n \int_0^t G(u)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_j(u)}\mathbf{g}(u)I(C_i \geq u)dM_i^1(u).$$

$\mathbf{M}^g(t)$ satisfies the condition for Lemma A.2(i). Hence, we have $\sup_{t\in[0,t^*]}\|\mathbf{M}^g(t)\|_{\max} = O_p\left(\sqrt{\log(q')/n}\right)$. Also define

$$\widetilde{\Delta}_i(t) = \{\widetilde{\omega}_i(t) - I(C_i > t)\}Y_i(t).$$

By Lemma B.6, $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}|\Delta_i(t) - \widetilde{\Delta}_i(t)| = O_p\left(n^{-1/2}\right)$. The total variation of each $\Delta_i(t)$ is at most 2. Hence, we can apply integration by parts to (B.7),

$$\begin{aligned}
&G^{-1}(t^*)\mathbf{M}^g(t^*-)\otimes n^{-1/2}\sum_{j=1}^n \Delta_j(t^*-)\boldsymbol{\phi}(\mathbf{Z}_j(t^*-)) - n^{-1/2}\sum_{j=1}^n\int_0^{t^*}\mathbf{M}^g(t)\otimes\boldsymbol{\phi}(\mathbf{Z}_j(t))dM_j^{\Delta}(t)\\
&- n^{1/2}\int_0^{t^*}\mathbf{M}^g(t)\otimes G^{-1}(t)n^{-1}\sum_{j=1}^n\Delta_j(t)d\boldsymbol{\phi}(\mathbf{Z}_j(t))\\
&\triangleq I_1 - I_2 - I_3.
\end{aligned}$$

We have shown that $|\mathbf{M}^g(t^*-)| = O_p\left(\sqrt{\log(q')/n}\right)$ and $\sup_{j=1,\ldots,n}|\Delta_j(t^*-) - \widetilde{\Delta}_j(t^*-)| = O_p\left(n^{-1/2}\right)$. By assumption, $\|\boldsymbol{\phi}(\mathbf{Z}_j(t^*-))\|_{\max} \leq K_\phi \asymp 1$. As a result, we may replace the $\Delta_i(t)$ in I_1 by $\widetilde{\Delta}_i(t)$ with with an $O_p\left(\sqrt{\log(q')/n}\right)$ error. Since $\widetilde{\Delta}_j(t^*-)\boldsymbol{\phi}(\mathbf{Z}_j(t^*-))$'s are i.i.d. mean zero random variables, $\|n^{-1}\sum_{j=1}^n\widetilde{\Delta}_j(t^*-)\boldsymbol{\phi}(\mathbf{Z}_j(t^*-))\|_{\max} = O_p\left(\sqrt{\log(q)/n}\right)$ by Hoeffding (1963). Multiplying the rates together, we get $\|I_1\|_{\max} = O_p\left(\sqrt{\log(q)\log(q')/n}\right) = o_p(1)$.

I_2 can be expanded as

$$n^{-1/2}\sum_{j=1}^n\int_0^{t^*} G(t)^{-1}\mathbf{M}^g(t)\left\{I(C_j \geq t)\upsilon_j(t)h(\mathbf{Z}_j(t)) - \frac{\sum_{k=1}^n\omega_k(t-)\upsilon_k(t)h(\mathbf{Z}_k(t))}{\sum_{k=1}^n I(X_k \geq t)}I(X_j \geq t)\right\}dM_j^c(t)$$

By Lemma B.4*, $n\left\{\sum_{k=1}^n I(X_k \geq t)\right\}^{-1} = O_p(1)$. The integrand in I_2 is the product of $\mathbf{M}^g(t)$ and a $O_p(1)$ term. Hence, we can apply Lemma A.2(ii) to get $\|I_2\|_{\max} = O_p\left(\sqrt{\log(q')\log(qq')/n}\right) = o_p(1)$.

By (D3), we may further expand I_3 into

$$\begin{aligned}
& n^{1/2}\int_0^{t^*}\mathbf{M}^g(t)\otimes G^{-1}(t)n^{-1}\sum_{j=1}^n\Delta_j(t)\nabla\boldsymbol{\phi}(\mathbf{Z}_j(t))^\top\mathbf{d}_j^z(t)dt \\
& +n^{1/2}\int_0^{t^*}\mathbf{M}^g(t)\otimes G^{-1}(t)n^{-1}\sum_{j=1}^n\Delta_j(t)\triangle\boldsymbol{\phi}(\mathbf{Z}_j(t))dN_j^z(t) \\
\triangleq & I_3'+I_3'',
\end{aligned}$$

where $\triangle\boldsymbol{\phi}(\mathbf{Z}_j(t))=\boldsymbol{\phi}(\mathbf{Z}_j(t))-\boldsymbol{\phi}(\mathbf{Z}_j(t-))$. By assumption on $h(\mathbf{Z})$ and (D3), $|\nabla\boldsymbol{\phi}(\mathbf{Z}_j(t))^\top\mathbf{d}_j^z(t)|$ and $\triangle\boldsymbol{\phi}(\mathbf{Z}_j(t))$ are bounded by L_hL_z and L_hK, respectively. With $\sup_{t\in[0,t^*]}|\mathbf{M}^g(t)|=O_p\left(\sqrt{\log(q')/n}\right)$ and $N_j^z(t^*)<K_z=o(\sqrt{n/(\log(p)\log(n))})$, we may replace the $\Delta_j(t)$'s by $\widetilde{\Delta}_j(t)$'s with an $o_p(1)$ error. Each $\widetilde{\Delta}_j(t)\nabla\boldsymbol{\phi}(\mathbf{Z}_j(t))^\top\mathbf{d}_j^z(t)$ has mean zero and at most K_z+1 jumps, and it is $(L_hL_z+K_\phi L_{dz})$-Lipschitz between two consecutive jumps under (D3) and conditions on $\boldsymbol{\phi}(\mathbf{z})$. By applying Lemma A.1(ii), we get

$$\sup_{t\in[0,t^*]}\left\|n^{-1}\sum_{j=1}^n\widetilde{\Delta}_j(t)\nabla\boldsymbol{\phi}(\mathbf{Z}_j(t))^\top\mathbf{d}_j^z(t)\right\|_{\max}=O_p(\sqrt{\log(nq)/n}).$$

Hence, $\|I_3'\|_{\max}=O_p(\sqrt{\log(q')\log(nq)/n})+o_p(1)=o_p(1)$. By applying Lemma A.1(i) to

$$\{\widetilde{\Delta}_j(t)\triangle h(\mathbf{Z}_j(t)),N_j^z(t):j=1,\ldots,n\},$$

we get at the jumps of $N_i^z(t)$'s, at the t_{ik}, satisfy

$$\sup_{i=1,\ldots,n}\sup_{k\in 1\ldots K_N}\left|n^{-1}\sum_{j=1}^n\Delta_j(t_{ik})\triangle\boldsymbol{\phi}(\mathbf{Z}_j(t_{ik}))\right|=O_p(\sqrt{\log(nK_zq)/n})=O_p(\sqrt{\log(nq)/n}).$$

Hence, $\|I_3''\|_{\max}=O_p\left(K_z\sqrt{\log(nq)\log(q')/n}\right)=o_p(1)$. This completes the proof.

(iii) Define $\boldsymbol{\beta}_r=\boldsymbol{\beta}^o+r\{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\}$ and $h_j(r;t)=\bar{\mathbf{Z}}_j(t,\boldsymbol{\beta}_r)$. The subscript j means the j-th element of corespondent vector. By mean-value theorem, we have some $r\in(0,1)$ such that

$$\begin{aligned}
h_j(1;t)-h_j(0;t) &= \left(\{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\}^\top\frac{\mathbf{S}^{(2)}(t,\boldsymbol{\beta}_r)S^{(0)}(t,\boldsymbol{\beta}_r)-\mathbf{S}^{(1)}(t,\boldsymbol{\beta}_r)^{\otimes 2}}{S^{(0)}(t,\boldsymbol{\beta}_r)^2}\right)_j \\
&= \left(\{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\}^\top\sum_{i=1}^n\frac{\omega_i(t)Y_i(t)e^{\boldsymbol{\beta}_r^\top\mathbf{Z}_i(t)}}{nS^{(0)}(t,\boldsymbol{\beta}_r)}\{\mathbf{Z}_i(t)-\bar{\mathbf{Z}}_j(t,\boldsymbol{\beta}_r)\}^{\otimes 2}\right)_j
\end{aligned}$$

Since each $\left\|\{\mathbf{Z}_i(t)-\bar{\mathbf{Z}}_j(t,\boldsymbol{\beta}_r)\}^{\otimes 2}\right\|_{\max}\le K^2$ under (D2), their weighted average

$$\left\|\sum_{i=1}^n\frac{\omega_i(t)Y_i(t)e^{\boldsymbol{\beta}_r^\top\mathbf{Z}_i(t)}}{nS^{(0)}(t,\boldsymbol{\beta}_r)}\{\mathbf{Z}_i(t)-\bar{\mathbf{Z}}_j(t,\boldsymbol{\beta}_r)\}^{\otimes 2}\right\|_{\max}\le K^2.$$

Hence, we have shown that

$$\sup_{t\in[0,t^*]}\|\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)-\bar{\mathbf{Z}}(t,\widetilde{\boldsymbol{\beta}})\|_\infty \le \|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1 K^2 = O_p(\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1).$$

By a similar argument, we can show for some $r\in(0,1)$

$$\frac{e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}}{S^{(0)}(t,\boldsymbol{\beta}^o)}-\frac{e^{\widetilde{\boldsymbol{\beta}}^\top\mathbf{Z}_i(t)}}{S^{(0)}(t,\widetilde{\boldsymbol{\beta}})}=\frac{e^{\boldsymbol{\beta}_r^\top\mathbf{Z}_i(t)}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o)^\top}{S^{(0)}(t,\boldsymbol{\beta}_r)}\sum_{j=1}^n\frac{\omega_j(t)Y_j(t)e^{\boldsymbol{\beta}_r^\top\mathbf{Z}_j(t)}}{nS^{(0)}(t,\boldsymbol{\beta}_r)}\{\mathbf{Z}_i(t)-\mathbf{Z}_j(t)\}.$$

On event $\{\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1\le K_b,\ n^{-1}\sum_{i=1}^n I(X_i\ge t^*)\ge r_*/2\}$, we have

$$\inf_{t\in[0,t^*]}S^{(0)}(t,\widetilde{\boldsymbol{\beta}})>r^*/2*e^{2K_b},\quad \inf_{t\in[0,t^*]}S^{(0)}(t,\boldsymbol{\beta}^o)>r^*/2*e^{K_b}.$$

Hence,

$$|e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}/S^{(0)}(t,\boldsymbol{\beta}^o)-e^{\widetilde{\boldsymbol{\beta}}^\top\mathbf{Z}_i(t)}/S^{(0)}(t,\widetilde{\boldsymbol{\beta}})|\le\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1 2Ke^{4K_b}/r_*=O_p\left(\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1\right).$$

The event occurs with probability tending to one because we have$\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1=o_p(1)$ from Theorem 1* and $\sup_{t\in[0,t^*]}|S^{(0)}(t,\boldsymbol{\beta}^o)^{-1}|=O_p(1)$ from Lemma B.4*.

□

Proof of Lemma B.9.

To simplify notation, wherever possible we will use $\widehat{\Gamma}_j(\boldsymbol{\gamma})=\Gamma_j(\boldsymbol{\gamma},\widehat{\boldsymbol{\beta}})$.

(i) We want to prove that for all $j=1,\dots,p$, the differences $\widetilde{\boldsymbol{\gamma}}_j:=\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j^*$ belong to a certain convex cone.

It follows from the KKT conditions that, for $l=1,\dots,p-1$,

$$\begin{cases}\frac{\partial\widehat{\Gamma}_j(\widehat{\boldsymbol{\gamma}}_j)}{\partial\boldsymbol{\gamma}_{j,l}}+\lambda_j\mathrm{sgn}(\widehat{\boldsymbol{\gamma}}_{j,l})=0 & \text{if } \widehat{\boldsymbol{\gamma}}_{j,l}\neq 0;\\ \left|\frac{\partial\widehat{\Gamma}_j(\widehat{\boldsymbol{\gamma}}_j)}{\partial\boldsymbol{\gamma}_{j,l}}\right|\le\lambda_j & \text{if } \widehat{\boldsymbol{\gamma}}_{j,l}=0.\end{cases}$$

Denote $\mathcal{O}_j:=\left\{l\in\{1,\dots,p-1\}:\boldsymbol{\gamma}_{j,l}^*\neq 0\right\}$ and $\mathcal{O}_j^c:=\{1,\dots,p-1\}\setminus\mathcal{O}_j$. For $\xi_j>1$, it follows from the KKT conditions above that on the event

$$\Omega_0:=\{\|\nabla_\gamma\widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*)\|_\infty\le(\xi_j-1)\lambda_j/(\xi_j+1)\},$$

with $\bar{\boldsymbol{\gamma}}_j=\alpha\widehat{\boldsymbol{\gamma}}_j+(1-\alpha)\boldsymbol{\gamma}_j^*$, $\alpha\in(0,1)$

$$\begin{aligned}0&\le 2\widetilde{\boldsymbol{\gamma}}_j^\top\nabla_\gamma^2\widehat{\boldsymbol{\Gamma}}_j(\bar{\boldsymbol{\gamma}}_j)\widetilde{\boldsymbol{\gamma}}_j\\&=\widetilde{\boldsymbol{\gamma}}_j^\top\left\{\nabla_\gamma\widehat{\boldsymbol{\Gamma}}_j(\widehat{\boldsymbol{\gamma}}_j)-\nabla_\gamma\widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*)\right\}\\&=\sum_{l\in\mathcal{O}_j^c}\widetilde{\boldsymbol{\gamma}}_{j,l}\frac{\partial\widehat{\Gamma}_j(\widehat{\boldsymbol{\gamma}}_j)}{\partial\boldsymbol{\gamma}_{j,l}}+\sum_{l\in\mathcal{O}_j}\widetilde{\boldsymbol{\gamma}}_{j,l}\frac{\partial\widehat{\Gamma}_j(\widehat{\boldsymbol{\gamma}}_j)}{\partial\boldsymbol{\gamma}_{j,l}}-\widetilde{\boldsymbol{\gamma}}_j^\top\nabla_\gamma\widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*)\\&\le-\lambda_j\sum_{l\in\mathcal{O}_j^c}\widehat{\boldsymbol{\gamma}}_{j,l}\mathrm{sgn}(\widehat{\boldsymbol{\gamma}}_{j,l})+\lambda_j\sum_{l\in\mathcal{O}_j}|\widetilde{\boldsymbol{\gamma}}_{j,l}|+\frac{(\xi_j-1)\lambda_j}{\xi_j+1}\|\widetilde{\boldsymbol{\gamma}}_{j,\mathcal{O}_j}\|_1+\frac{(\xi_j-1)\lambda_j}{\xi_j+1}\|\widetilde{\boldsymbol{\gamma}}_{j,\mathcal{O}_j^c}\|_1\\&=-\frac{2\lambda_j}{\xi_j+1}\|\widetilde{\boldsymbol{\gamma}}_{j,\mathcal{O}_j^c}\|_1+\frac{2\xi_j\lambda_j}{\xi_j+1}\|\widetilde{\boldsymbol{\gamma}}_{j,\mathcal{O}_j}\|_1.\end{aligned}$$

(ii) Let $\mathbf{v} = \widetilde{\boldsymbol{\gamma}}/\|\widetilde{\boldsymbol{\gamma}}\|_1$ be the l_1-standardized direction for $\widetilde{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$. By part (i) and convexity of Γ_j in $\boldsymbol{\gamma}_j$, any $x \in (0, \|\widetilde{\boldsymbol{\gamma}}\|_1]$ satisfies

$$\mathbf{v}^\top \left\{ \nabla_\gamma \widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}^* + x\mathbf{v}) - \nabla_\gamma \widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}^*) \right\} \leq -\frac{2\lambda_j}{\xi_j + 1} \|\mathbf{v}_{\mathcal{O}_j^c}\|_1 + \frac{2\xi_j \lambda_j}{\xi_j + 1} \|\mathbf{v}_{\mathcal{O}_j}\|_1.$$

We relax the inequality about x above to establish an upper bound for $\|\widetilde{\boldsymbol{\gamma}}\|_1$. By the definition of κ_j, the left hand side can be bounded by

$$\mathbf{v}^\top \left\{ \nabla_\gamma \widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}^* + x\mathbf{v}) - \nabla_\gamma \widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}^*) \right\} = x\mathbf{v}^\top \nabla_\gamma^2 \widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}^*)\mathbf{v} \geq \frac{x\|\mathbf{v}_{\mathcal{O}_j}\|_1^2 \kappa_j(\xi_j, \mathcal{O}_j)}{s_j}.$$

The right hand side can be bounded using the complete square $\{\|\mathbf{v}_{\mathcal{O}_j}\|_1 - 2/(\xi_j + 1)\}^2$,

$$-\frac{2\lambda_j}{\xi_j + 1} \|\mathbf{v}_{\mathcal{O}_j^c}\|_1 + \frac{2\xi_j \lambda_j}{\xi_j + 1} \|\mathbf{v}_{\mathcal{O}_j}\|_1 = 2\lambda_j \|\mathbf{v}_{\mathcal{O}_j}\|_1 - \frac{2\lambda_j}{\xi_j + 1} \leq \lambda_j(\xi_j + 1)\|\mathbf{v}_{\mathcal{O}_j}\|_1^2.$$

Combining the bounds for both sides in the inequality, we get an upper bound for $\|\widetilde{\boldsymbol{\gamma}}\|_1$.

□

Proof of Lemma B.10.

We define

$$\widetilde{\Gamma}_j(\boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^n \int_0^{t^*} \{Z_{ij}(t) - \mu_j(t) - \boldsymbol{\gamma}^\top \mathbf{Z}_{i,-j}(t) + \boldsymbol{\gamma}^\top \boldsymbol{\mu}_{-j}(t)\}^2 dN_i^o(t).$$

By Lemmas B.8 and 3, $\max_{j=1,\ldots,p} \|\nabla_\gamma \widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*, \widehat{\boldsymbol{\beta}}) - \nabla_\gamma \widetilde{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*)\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right)$. $\nabla_\gamma \widetilde{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*)$ is the average of i.i.d. vectors with mean $\nabla_\gamma \bar{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*) = \mathbf{0}$ and maximal bound $K^2(1 + K_\gamma)$. We can apply Hoeffding (1963) to the matrix $(\nabla_\gamma \widetilde{\boldsymbol{\Gamma}}_1(\boldsymbol{\gamma}_1^*), \ldots, \nabla_\gamma \widetilde{\boldsymbol{\Gamma}}_p(\boldsymbol{\gamma}_p^*))$ to get

$$\max_{j=1,\ldots,p} \|\nabla_\gamma \widetilde{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*)\|_\infty = \|(\nabla_\gamma \widetilde{\boldsymbol{\Gamma}}_1(\boldsymbol{\gamma}_1^*), \ldots, \nabla_\gamma \widetilde{\boldsymbol{\Gamma}}_p(\boldsymbol{\gamma}_p^*))\|_{\max} = O_p(\sqrt{\log(p^2)/n}) = O_p(\sqrt{\log(p)/n}).$$

□

Proof of Lemma B.11.

(i) We define

$$\widetilde{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^n \int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}^{\otimes 2} dN_i^o(t).$$

The total variation of each $N_i^o(t)$ is at most 1. By Lemma B.8, we have $\sup_{t \in [0,t^*]} \|\bar{\mathbf{Z}}(t, \widehat{\boldsymbol{\beta}}) - \boldsymbol{\mu}\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right)$. Hence,

$$\|\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}}\|_{\max} \leq 2KO_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right) = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right).$$

Now, $\widetilde{\boldsymbol{\Sigma}}$ is average of i.i.d. with mean $\boldsymbol{\Sigma}$ and bounded maximal norm K^2. We apply Hoeffding (1963) with union bound,

$$\Pr\left(\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \geq K^2 x\right) \leq 2p^2 e^{2nx^2}.$$

Choosing $x = \sqrt{\log(2p^2/\varepsilon)/(2n)}$, we have $\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = O_p(\sqrt{\log(p)/n})$.

(ii) We alternatively use the following form

$$\ddot{\mathbf{m}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left\{n^{-1}\sum_{j=1}^{n}\frac{\omega_j(t)Y_j(t)e^{\boldsymbol{\beta}^\top\mathbf{Z}_j(t)}}{S^{(0)}(t,\boldsymbol{\beta})}\mathbf{Z}_i(t)^{\otimes 2} - \bar{\mathbf{Z}}(t,\boldsymbol{\beta})^{\otimes 2}\right\}dN_i^o(t).$$

By Lemma B.8(iii), we have

$$\|\ddot{\mathbf{m}}(\widetilde{\boldsymbol{\beta}}) - \ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_{\max} = O_p(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1).$$

We also have a similar form for

$$\ddot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left\{n^{-1}\sum_{j=1}^{n}\frac{I(C_j\geq t)Y_j(t)e^{\boldsymbol{\beta}^\top\mathbf{Z}_j(t)}}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta})}\mathbf{Z}_i(t)^{\otimes 2} - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta})^{\otimes 2}\right\}dN_i^o(t).$$

By Lemma B.8(i), we have

$$\|\ddot{\mathbf{m}}(\boldsymbol{\beta}^o) - \ddot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)\|_{\max} = O_p\left(\sqrt{\log(p)/n}\right).$$

Finally, we use the martingale property of

$$\begin{aligned}\ddot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o) - \widetilde{\boldsymbol{\Sigma}} =& n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left\{\frac{\widetilde{\mathbf{S}}^{(2)}(t,\boldsymbol{\beta}^o)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)} - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)^{\otimes 2}\right\}I(C_i\geq t)dM_i^1(t)\\ &- n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_i(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}^{\otimes 2}I(C_i\geq t)dM_i^1(t)\\ &+ n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left[\{\mathbf{Z}_i(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}^{\otimes 2} - \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}^{\otimes 2}\right]I(C_i\geq t)dN_i^o(t)\end{aligned}$$

under filtration $\mathcal{F}_t^*$. The integrands in the first two martingale terms are bounded by K^2. Hence, we can apply Lemma A.2(ii) to obtain that their maximal norms are both $O_p\left(\sqrt{\log(p)/n}\right)$. We apply Lemma B.8(i) to the integrand of the third term, equivalently expressed as

$$\{\boldsymbol{\mu}(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}\{\mathbf{Z}_i(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}^\top + \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\{\boldsymbol{\mu}(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}^\top.$$

Therefore, we obtain $\|\ddot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o) - \widetilde{\boldsymbol{\Sigma}}\|_{\max} = O_p\left(\sqrt{\log(p)/n}\right)$.

We put the rates together by the triangle inequality.

□

Proof of Lemma B.12.

The proof is similar to that of Lemma 2. Define the compatibility factor for $\mathcal{C}_j(\xi_j,\mathcal{O}_j)$ and symmetric matrix $\boldsymbol{\Phi}$ as

$$\kappa_j(\xi_j,\mathcal{O}_j;\boldsymbol{\Phi}) = \sup_{0\neq\mathbf{g}\in\mathcal{C}_j(\xi_j,\mathcal{O}_j)}\frac{\sqrt{s_j\mathbf{g}^\top\boldsymbol{\Phi}\mathbf{g}}}{\|\mathbf{g}_{\mathcal{O}_j}\|_1}.$$

Apparently, $\kappa_j(\xi_j, \mathcal{O}_j) = \kappa_j\big(\xi_j, \mathcal{O}_j; \nabla^2_\gamma \mathbf{\Gamma}(\boldsymbol{\gamma}^*, \widehat{\boldsymbol{\beta}})\big)$. Notice that

$$\nabla^2_\gamma \mathbf{\Gamma}(\boldsymbol{\gamma}^*, \widehat{\boldsymbol{\beta}}) = n^{-1} \sum_{i=1}^{n} \int_0^{t^*} \{\mathbf{Z}_{i,-j}(t) - \bar{\mathbf{Z}}_{-j}(t, \widehat{\boldsymbol{\beta}})\}^{\otimes 2} dN_i^o(t) = \widehat{\mathbf{\Sigma}}_{-j,-j},$$

where $\widehat{\mathbf{\Sigma}}_{-j,-j}$ is a $\widehat{\mathbf{\Sigma}}$ dropping its jth row and column. By Lemma 4.1 in Huang *et al.* (2013) (for a similar result, see van de Geer and Bühlmann (2009) Corollary 10.1),

$$\kappa_j(\xi_j, \mathcal{O}_j)^2 = \kappa_j^2\big(\xi_j, \mathcal{O}_j; \widehat{\mathbf{\Sigma}}_{-j,-j}\big) \geq \kappa_j^2(\xi_j, \mathcal{O}_j; \mathbf{\Sigma}_{-j,-j}) - s_j(\xi_j + 1)^2 \|\mathbf{\Sigma}_{-j,-j} - \widehat{\mathbf{\Sigma}}_{-j,-j}\|_{\max}.$$

For any non-zero $\mathbf{g} \in \mathbb{R}^{p-1}$, let $\mathbf{g}^*$ be its embedding into $\mathbb{R}^p$ defined as

$$g_k^* = \begin{cases} g_k & k < j \\ 0 & k = j \\ g_{k-1} & k > j \end{cases}$$

Then, we may establish a lower bound for the smallest eigenvalue of $\mathbf{\Sigma}_{-j,-j}$ by (D4)

$$\inf_{0 \neq \mathbf{g} \in \mathbb{R}^{p-1}} \mathbf{g}^\top \mathbf{\Sigma}_{-j,-j} \mathbf{g} = \inf_{0 \neq \mathbf{g} \in \mathbb{R}^{p-1}} \mathbf{g}^{*\top} \mathbf{\Sigma} \mathbf{g}^* \geq \rho_* \|\mathbf{g}\|_2^2.$$

Hence, $\inf_{j=1,\ldots,p} \kappa_j^2(\xi_j, \mathcal{O}_j; \mathbf{\Sigma}_{-j,-j}) \geq \rho_*$. Using the result in Lemma B.11(i) under (D5), we have

$$\inf_{j=1,\ldots,p} \kappa_j(\xi_j, \mathcal{O}_j)^2 \geq \rho_* - \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_{\max} s_{\max} \max_{j=1,\ldots,p} (\xi_j + 1)^2 = \rho_* - o_p(1).$$

Therefor, if $\xi_{\max} \asymp 1$, we must have that $\big\{\inf_j \kappa_j(\xi_j, \mathcal{O}_j)^2 \geq \rho_*/2\big\}$ occurs with probability tending to one. □

References

Andersen, P. K. and Gill, R. (**1982**). "Cox's regression model for counting processes: a large sample study", Annals of Statistics **10**, 1100–1120.

Azuma, K. (**1967**). "Weighted sums of cerntain dependent random variables", Tôhoku Mathematical Journal **19**, 357–367.

Basu, S. and Michailidis, G. (**2015**). "Regularized estimation in sparse high-dimensional time series models", The Annals of Statistics **43**, 1535–1567.

Belloni, A. and Chernozhukov, V. (**2011**). "l1-penalized quantile regression in high-dimensional sparse models", The Annals of Statistics **39**, 82–130.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (**2009**). "Simultaneous analysis of LASSO and Dantzig selector", Annals of Statistics **37**, 1705–1732.

Binder, H., Allignol, A., Schumacher, M., and Beyersmann, J. (**2009**). "Boosting for high-dimensional time-to-event data with competing risks", Bioinformatics **25**, 890–896.

Bradic, J., Fan, J., and Jiang, J. (**2011**). "Regularization for cox's proportional hazards model with np-dimensionality", Annals of Statistics **39**, 3092.

Bradic, J. and Song, R. (**2015**). "Structured estimation for the nonparametric cox model", Electronic Journal of Statistics **9**, 492–534.

Bühlmann, P. and Van De Geer, S. (**2011**). *Statistics for high-dimensional data: methods, theory and applications* (Springer Science & Business Media).

Cho, H. and Fryzlewicz, P. (**2015**). "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **77**, 475–507.

Durrett, R. (**2013**). *Probability: Theory and Examples, 4th edition* (Cambridge University Press).

Fan, J. and Li, R. (**2001**). "Variable selection via nonconcave penalized likelihood and its oracle properties", Journal of the American statistical Association **96**, 1348–1360.

Fan, J. and Lv, J. (**2010**). "A selective overview of variable selection in high dimensional feature space", Statistica Sinica **20**, 101.

Fang, E. X., Ning, Y., and Liu, H. (**2017**). "Testing and confidence intervals for high dimensional proportional hazards models", to appear in Journal of the Royal Statistical Society: Series B (Statistical Methodology) .

Fine, J. P. and Gray, R. J. (**1999**). "A proportional hazard model for the subdistribution of a competing risk", Journal of the American Statistical Association **94**, 496–509.

Gaïffas, S. and Guilloux, A. (**2012**). "High-dimensional additive hazards models and the lasso", Electronic Journal of Statistics **6**, 522–546.

Hoeffding, W. (**1963**). "Probability inequalities for sums of bounded random variables", Journal of the American Statistical Association **58**, 13–30.

Hou, J., Paravati, A., Xu, R., and Murphy, J. (**2017**). "High-Dimensional Variable Selection and Prediction under Competing Risks with Application to SEER-Medicare Linked Data", ArXiv e-prints:1704.07989 .

Huang, J., Ma, S., and Xie, H. (**2006**). "Regularized estimation in the accelerated failure time model with high-dimensional covariates", Biometrics **62**, 813–820.

Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C.-H. (**2013**). "Oracle inequalities for the LASSO in the Cox model", Annals of Statistics **41**, 1142–1165.

Johnson, B. A. (**2008**). "Variable selection in semiparametric linear regression with censored data", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**, 351–370.

Kalbfleisch, J. D. and Prentice, R. L. (**2002**). *The Statistical Analysis of Failure Time Data (2nd ed.)* (John Wiley & Sons, Inc., Hoboken, New Jersey).

Lemler, S. (**2013**). "Oracle inequalities for the lasso in the high-dimensional multiplicative aalen intensity model", Les Annales de l'Institut Henri Poincaré, arXiv preprint **21**, 109–137.

Meinshausen, N. and Bühlmann, P. (**2006**). "High-dimensional graphs and variable selection with the lasso", The Annals of Statistics 1436–1462.

Meinshausen, N. and Yu, B. (**2009**). "Lasso-type recovery of sparse representations for high-dimensional data", The Annals of Statistics 246–270.

Murphy, S. A. (**1994**). "Consistency in a proportional hazards model incorporating a random effect", Annals of Statistics **22**, 712–731.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (**2011**). "Support union recovery in high-dimensional multivariate regression", The Annals of Statistics **39**, 1–47.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (**2010**). "High-dimensional ising model selection using l1-regularized logistic regression", The Annals of Statistics **38**, 1287–1319.

Sun, H., Lin, W., Feng, R., and Li, H. (**2014**). "Network-regularized high-dimensional cox regression for analysis of genomic data", Statistica Sinica **24**, 1433.

Tibshirani, R. (**1996**). "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society. Series B (Methodological) 267–288.

van de Geer, S. and Bühlmann, P. (**2009**). "On the conditions used to prove aracle results for the Lasso", Electronic Journal of Statistics **3**, 1360–1392.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (**2014**). "On asymptotically optimal confidence regions and tests for high-dimensional models", Annals of Statistics **42**, 1166–1202.

van de Geer, S. A. (**2007**). *The deterministic LASSO.*, Technical Report 140 (ETH Zürich, Switzerland).

Wasserman, L. and Roeder, K. (**2009**). "High dimensional variable selection", The Annals of Statistics **37**, 2178.

Zhang, C.-H. and Zhang, S. S. (**2014**). "Confidence intervals for low dimensional parameters in high dimensional linear models", Journal of the Royal Statistical Society, Series B **76**, 217–242.

Zhou, S., Rütimann, P., Xu, M., and Bühlmann, P. (**2011**). "High-dimensional covariance estimation based on gaussian graphical models", Journal of Machine Learning Research **12**, 2975–3026.